

2017-12-19

Cross-Situational Learning with Bayesian Generative Models for Multimodal Category and Word Learning in Robots

Taniguchi, A

<http://hdl.handle.net/10026.1/111192>

10.3389/fnbot.2017.00066

FRONTIERS IN NEUROBOTICS

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.



Cross-Situational Learning with Bayesian Generative Models for Multimodal Category and Word Learning in Robots

Akira Taniguchi^{1*}, Tadahiro Taniguchi¹ and Angelo Cangelosi²

¹Emergent Systems Laboratory, Ritsumeikan University, Kusatsu, Japan ²The Centre for Robotics and Neural Systems, Plymouth University, Plymouth, United Kingdom

In this paper, we propose a Bayesian generative model that can form multiple categories based on each sensory-channel and can associate words with any of the four sensory-channels (action, position, object, and color). This paper focuses on cross-situational learning using the co-occurrence between words and information of sensory-channels in complex situations rather than conventional situations of cross-situational learning. We conducted a learning scenario using a simulator and a real humanoid iCub robot. In the scenario, a human tutor provided a sentence that describes an object of visual attention and an accompanying action to the robot. The scenario was set as follows: the number of words per sensory-channel was three or four, and the number of trials for learning was 20 and 40 for the simulator and 25 and 40 for the real robot. The experimental results showed that the proposed method was able to estimate the multiple categorizations and to learn the relationships between multiple sensory-channels and words accurately. In addition, we conducted an action generation task and an action description task based on word meanings learned in the cross-situational learning scenario. The experimental results showed that the robot could successfully use the word meanings learned by using the proposed method.

OPEN ACCESS

Edited by:

Frank Van Der Velde,
University of Twente, Netherlands

Reviewed by:

Maxime Petit,
Imperial College London,
United Kingdom
Yulia Sandamirskaya,
University of Zurich, Switzerland

*Correspondence:

Akira Taniguchi
a.taniguchi@em.ci.ritsumei.ac.jp

Received: 17 July 2017

Accepted: 21 November 2017

Published: 19 December 2017

Citation:

Taniguchi A, Taniguchi T and
Cangelosi A (2017) Cross-Situational
Learning with Bayesian Generative
Models for Multimodal Category and
Word Learning in Robots.
Front. Neurobot. 11:66.
doi: 10.3389/fnbot.2017.00066

Keywords: Bayesian model, cross-situational learning, lexical acquisition, multimodal categorization, symbol grounding, word meaning

1. INTRODUCTION

This paper addresses the study of robotic learning of the word meanings inspired by the process of language acquisition of humans. We developed an unsupervised machine learning method to enable linguistic interaction between humans and robots. Human infants can acquire word meanings by estimating the relationships between multimodal information and words in a variety of situations. For example, if an infant grasps a green cup by hand, let us consider the way the parent describes the actions of the infant to the infant using a sentence such as “grasp green front cup.” In this case, the infant does not know the relationship between words and situations because it has not acquired the meanings of words. In other words, the infant cannot determine whether the word “green” indicates an action, an object, or a color. However, it is believed that the infant can learn that the word “green” represents the green color by observing the co-occurrence of the word “green” with objects of green color in various situations. This is known as cross-situational learning (CSL), which has been both studied in children (Smith et al., 2011) and modeled in simulated agents and robots (Fontanari et al., 2009). The CSL is related to the symbol grounding problem (Harnad, 1990), which is a challenging and significant issue in robotics.

The generalization ability and the robustness of observation noise to process situations that have never been experienced are important in cognitive robotics. The study of language acquisition by infants led to the proposal of a hypothesis of taxonomic bias (Markman and Hutchinson, 1984) that infants tend to understand a word as the name of a category to which the target object belongs rather than a proper noun. This hypothesis could also be considered to play an important role in CSL. In this study, we assume that words are associated with categories based on taxonomic bias. By associating words with categories, it becomes possible for a human to generalize and process words. Therefore, humans can use words for communication in new situations. To develop this ability, the robot needs to form categories from observation information autonomously. We develop this ability by categorization based on the Bayesian generative model. Another hypothesis regarding the lexical acquisition by an infant was mutual exclusivity bias (constraint) (Markman and Wachtel, 1988). In studies on lexical acquisition, this hypothesis was considered to be particularly important for CSL (Twomey et al., 2016). Mutual exclusivity bias assumes that the infant considers the name of an object to correspond to one particular category only. In other words, multiple categories do not correspond to that word simultaneously. In Imai and Mazuka (2007), it was suggested that once an infant decides whether a word refers to the name of an object or a substance, the same word is not applied across the ontological distinction such as objects and substances. In this study, we extend the mutual exclusivity constraint to the CSL problem in complex situations. We aim to develop a novel method that can acquire knowledge of multiple categories and word meanings simultaneously. In addition, we verify whether the effect of mutual exclusivity is biased toward lexical acquisition by constructing a model assuming different constraints.

In addition, humans can perform the instructed action using acquired knowledge. For example, the parent places some objects in front of an infant and speaks “grasp green right ball” to the infant. In this case, the infant can use the acquired word meanings to select the green ball to the right of some objects and perform the action of grasping. Furthermore, humans can explain self-action with the sentence using the acquired knowledge. For example, if the infant knows the word meanings after grasping a blue box in front of it, the infant can speak “grasp blue front box” to another person. Understanding instructions and describing situations are crucial problems that are also required to build a cognitive robot.

In this paper, the goal is to develop an unsupervised machine-learning method for learning the relationships between words and the four sensory-channels (action, object, color, and position) from the robot's experience of observed sentences describing object manipulation scenes. In the above example, sentences containing four words for four sensory-channels are shown. However, in the scenario described in this study, sentences of less than four words are allowed. In addition, the position sensory-channel corresponds to the original position of the object. In other words, we assume that the environment is static. We assume that the robot can recognize spoken words without errors, as this work focuses specifically on (1) the categorization for each sensory-channel, (2) the learning of relationships between words and sensory-channels, and (3) the grounding of words in multiple categories.

In addition, we demonstrate whether the robot can carry out its actions and the sentence description of its action by conducting experiments using the CSL results. The main contributions of this paper are as follows:

- We proposed an unsupervised machine-learning method based on a Bayesian generative model that makes it possible to learn word meanings, i.e., the relationships between words and categories, from complex situations.
- We demonstrated that word meanings learned by using the proposed method are effective for generating an action and description of a situation.

The remainder of this paper is organized as follows. In Section 2, we discuss previous studies on lexical acquisition by a robot and CSL that are relevant to our study. In Section 3, we present a proposed Bayesian generative model for CSL. In Sections 4 and 5, we discuss the effectiveness of the proposed method in terms of three tasks, i.e., cross-situational learning, action generation, and an action description task, in a simulation and a real environment, respectively. Section 6 concludes the paper.

2. RELATED WORK

2.1. Lexical Acquisition by Robot

Studies of language acquisition also constitute a constructive approach to the human developmental process (Cangelosi and Schlesinger, 2015), the language grounding (Steels and Hild, 2012), and the symbol emergence (Taniguchi et al., 2016c). One approach to studying language acquisition focuses on the estimation of phonemes and words from speech signals (Goldwater et al., 2009; Heymann et al., 2014; Taniguchi et al., 2016d). However, these studies used only continuous speech signals without using co-occurrence based on other sensor information, e.g., visual, tactile, and proprioceptive information. Therefore, the robot was not required to understand the meaning of words. Yet, it is important for a robot to understand word meanings, i.e., grounding the meanings to words, for human–robot interaction (HRI).

Roy and Pentland (2002) proposed a computational model by which a robot could learn the names of objects from images of the object and natural infant-directed speech. Their model could perform speech segmentation, lexical acquisition, and visual categorization. Hörnstein et al. (2010) proposed a method based on pattern recognition and hierarchical clustering that mimics a human infant to enable a humanoid robot to acquire language. Their method allowed the robot to acquire phonemes and words from visual and auditory information through interaction with the human. Nakamura et al. (2011a,b) proposed multimodal latent Dirichlet allocation (MLDA) and a multimodal hierarchical Dirichlet process (MHDP) that enables the categorization of objects from multimodal information, i.e., visual, auditory, haptic, and word information. Their methods enabled more accurate object categorization by using multimodal information. Taniguchi et al. (2016a) proposed a method for simultaneous estimation of self-positions and words from noisy sensory information and an

uttered word. Their method integrated ambiguous speech recognition results with the self-localization method for learning spatial concepts. However, Taniguchi et al. (2016a) assumed that the name of a place would be learned from an uttered word. Taniguchi et al. (2016b) proposed a nonparametric Bayesian spatial concept acquisition method (SpCoA) based on place categorization and unsupervised word segmentation. SpCoA could acquire the names of places from spoken sentences including multiple words. In the above studies, the robot was taught to focus on one target, e.g., an object or a place, by a tutor using one word or one sentence. However, considering a more realistic problem, the robot needs to know which event in a complicated situation is associated with which word in the sentence. The CSL, which is extended from the aforementioned studies on the lexical acquisition, is a more difficult and important problem in robotics in comparison. Our research concerns the CSL problem because of its importance in relation to the lexical acquisition by a robot.

2.2. Cross-Situational Learning

2.2.1. Conventional Cross-Situational Learning Studies

Frank et al. (2007, 2009) proposed a Bayesian model that unifies statistical and intentional approaches to cross-situational word learning. They conducted basic CSL experiments with the purpose of teaching an object name. In addition, they discussed that the effectiveness of mutual exclusivity for CSL in probabilistic models. Fontanari et al. (2009) performed object-word mapping from the co-occurrence between objects and words by using a method based on neural modeling fields (NMF). In “modi” experiments using iCub, their findings were similar to those reported by Smith and Samuelson (2010). The abovementioned studies are CSL studies that were inspired by studies based on experiments with human infants. These studies assumed a simple situation such as learning the relationship between objects and words as the early stage of CSL. However, the real environment is varied and more complex. In this study, we focus on the problem of CSL in utterances including multiple words and observations from multiple sensory-channels.

2.2.2. Probabilistic Models

Qu and Chai (2008, 2010) proposed a learning method that automatically acquires novel words for an interactive system. They focused on the co-occurrence between word-sequences and entity-sequences tracked by eye-gaze in lexical acquisition. Qu and Chai’s method, which is based on the IBM-translation model (Brown et al., 1993), estimates the word-entity association probability. However, their studies did not result in perfect unsupervised lexical acquisition because they used domain knowledge based on WordNet. Matuszek et al. (2012) presented a joint model of language and perception for grounded attribute learning. This model enables the identification of which novel words correspond to color, shape, or no attribute at all. Celikkanat et al. (2014) proposed an unsupervised learning method based on latent Dirichlet allocation (LDA) that allows many-to-many relationships between objects and contexts. Their method was able to predict the context from the observation information and plan the action using learned contexts. Chen et al. (2016) proposed an active learning method for cross-situational learning

of object-word association. In experiments, they showed that LDA was more effective than non-negative matrix factorization (NMF). However, they did not perform any HRI experiment using the learned language. In our study, we perform experiments that use word meanings learned in CSL to generate an action and explain a current situation.

2.2.3. Neural Network Models

Yamada et al. (2015, 2016) proposed a learning method based on a stochastic continuous-time recurrent neural network (CTRNN) and a multiple time-scales recurrent neural network (MTRNN). They showed that the learned network formed an attractor structure representing both the relationships between words and action and the temporal pattern of the task. Stramandinoli et al. (2017) proposed partially recurrent neural networks (P-RNNs) for learning the relationships between motor primitives and objects. Zhong et al. (2017) proposed multiple time-scales gated recurrent units (MTGRU) inspired by MTRNN and long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997). They showed that the MTGRU could learn long-term dependencies in large-dimensional multimodal datasets by conducting multimodal interaction experiments using iCub. The learning results of the above studies using neural networks (NNs) are difficult to interpret because time-series data is mapped to continuous latent space. These studies implicitly associate words with objects and actions. Generally, NN methods require a massive amount of learning data in many cases. On the other hand, the learning result is easier to interpret when Bayesian methods rather than NN methods are used. In addition, Bayesian methods require less data to learn efficiently. We propose a Bayesian generative model that can perform CSL, including action learning.

2.2.4. Robot-to-Robot Interaction

Spranger (2015) and Spranger and Steels (2015) proposed a method for the co-acquisition of semantics and syntax in the spatial language. The experimental results showed that the robot could acquire spatial grammar and categories related to spatial direction. Heath et al. (2016) implemented mobile robots (Lingodroids) capable of learning a lexicon through robot-to-robot interaction. They used two robots equipped with different sensors and simultaneous localization and mapping (SLAM) algorithms. These studies reported that the robots created their lexicons in relation to places and the distance in terms of time. However, these studies did not consider lexical acquisition by HRI. We consider HRI to be necessary to enable a robot to learn human language.

2.2.5. Multimodal Categorization and Word Learning

Attamimi et al. (2016) proposed multilayered MLDA (mMLDA) that hierarchically integrates multiple MLDAs as an extension of Nakamura et al. (2011a). They performed an estimation of the relationships among words and multiple concepts by weighting the learned words according to their mutual information as a post-processing step. In their model, the same uttered words are generated from three kinds of concepts, i.e., this model has three variables for same word information in different concepts. We consider this to be an unnatural assumption as the generative model for generating words. However, in our proposed model, we assume that the uttered words are generated from one variable. We consider our proposed model to involve a more natural

assumption than Attamimi's model. In addition, their study did not use data that were autonomously obtained by the robot. In Attamimi et al. (2016), it was not possible for the robot to learn the relationships between self-actions and words because human motions obtained by the motion capture system based on Kinect and a wearable sensor device attached to a human were used as action data. In our study, the robot learns the action category based on subjective self-action. Therefore, the robot can perform a learned action based on a sentence of human speech. In this paper, we focus on complicated CSL problems arising from situations with multiple objects and sentences including words related to various sensory-channels such as the names, position, and color of objects, and the action carried out on the object.

3. MULTICHANNEL CATEGORIZATIONS AND LEARNING THE MEANING OF WORDS

We propose a Bayesian generative model for cross-situational learning. The proposed method can estimate categories of multiple sensory-channels and the relationships between words and sensory-channels simultaneously.

3.1. Overview of the Scenario and Assumptions

Here, we provide an overview of the scenario on which we focus and some of the assumptions in this study. **Figure 1** shows an overview of the scenario. The robot does not have any specific knowledge of objects, but it can recognize that objects exist on the table, i.e., the robot can segment the object and then extract the features of the segmented object. In addition, we assume that the robot can recognize the sentence uttered by the tutor without error. The training procedure consists of the following steps:

1. The robot is in front of the table on which the objects are placed. Multiple objects are placed separately on the table.
2. The robot selects an object from the objects on the table. The robot pays visual attention to a selected object, and then, exerts an action on the selected object, e.g., "grasp," "touch," "reach," and "look-at."
3. The human tutor utters a sentence including words about the object at which the robot is gazing and also a word about the action performed by the robot, e.g., "grasp front green cup."
4. The robot obtains multimodal information regarding all objects on the table in the current situation, e.g., the object features, positions, colors, and self-action. The robot processes the sentence to discover the meanings of the words.

This process (steps 1–4) is carried out many times in different situations.

We assume that the robot does not know the relationships between the words and sensory-channels in advance. This study does not consider grammar, i.e., a unigram language model is assumed. The robot learns word meanings and multiple categories by using visual, tactile, and proprioceptive information, as well as words.

In this study, we consider two-level cross-situational learning (CSL-I and II). The first level (CSL-I) is the selection of an object related to a tutor utterance from multiple objects on the table.

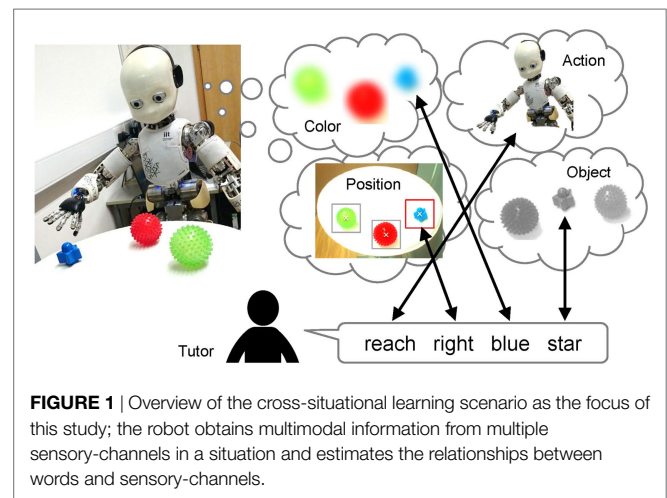


FIGURE 1 | Overview of the cross-situational learning scenario as the focus of this study; the robot obtains multimodal information from multiple sensory-channels in a situation and estimates the relationships between words and sensory-channels.

The second level (CSL-II) is the selection of the relationship between the specific word in the sentence and a sensory-channel in the multimodal information. In the first level, we assume joint attention. Tomasello and Farrar (1986) showed that the utterance referring to the object on which the child's attention was already focused is more effective in language acquisition. The above scenario enables the tutor to identify the object of attention, i.e., the object at which the robot is gazing. Furthermore, we assume that the robot considers the tutor to be speaking a sentence concerning the object of attention. This assumption of joint attention can avoid the problem of the selection of an object. The second level is the main problem in this study. Many previous studies on CSL-I have been reported (Frank et al., 2007, 2009; Fontanari et al., 2009; Morse et al., 2010); however, there are not the case for studies on CSL-II. The study discussed in this paper focused on solving the crucial problem of CSL-II.

In this study, we assume a two-level mutual exclusivity constraint (Markman and Wachtel, 1988) (MEC-I and II) regarding the selection of the sensory-channel. The first level (MEC-I) is the mutual exclusivity of sensory-channels with a word, i.e., one word is allocated to one category in one sensory-channel. The second level (MEC-II) is the mutual exclusivity between sensory-channels indicated by words, i.e., one word related to each sensory-channel is spoken only once in a sentence (or is not spoken). MEC-II is a stronger constraint than MEC-I. The proposed method can include both levels of mutual exclusivity.

3.2. Generative Model and Graphical Model

The generative model of the proposed method is defined as equations (1–10). **Figure 2** shows a graphical model representing the probabilistic dependencies between variables of the generative model. Basically, the categorization for each sensory-channel is based on the Gaussian mixture model (GMM). In this model, the probability distribution of words is represented by the categorical distribution. The categorization of words in sentences is similar to that of LDA. The latent variable of a word shares the latent variable of any one of the sensory-channels in GMMs, signifying that a word and a category in a particular sensory-channel are generated from the same latent variable.

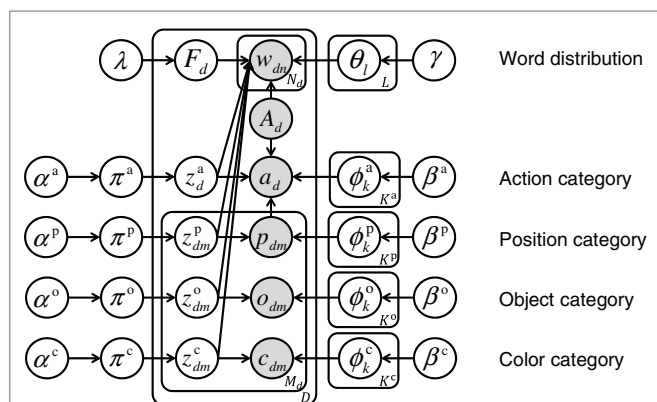


FIGURE 2 | Proposed graphical model for multichannel categorizations and for learning word meaning; the action, position, color, and object categories are represented by a component in Gaussian mixture models (GMMs). A word distribution is related to a category in GMMs. Gray nodes represent observed variables. Each variable is explained in the description of the generative model in Section 3.2.

We describe the generative model as follows:

$$F_d \sim \text{Unif}(\lambda) \quad (1)$$

$$\theta_l \sim \text{Dir}(\gamma) \quad (2)$$

$$\pi \sim \text{GEM}(\alpha) \quad (3)$$

$$z_{dm} \sim \text{Cat}(\pi) \quad (4)$$

$$w_{dn} \sim \text{Cat}\left(\theta_{l=(F_{dn}, z_{dm}^{F_{dn}})}\right) \quad (5)$$

$$\phi_k \sim \text{GIW}(\beta) \quad (6)$$

$$o_{dm} \sim \text{Gauss}(\phi_{z_{dm}^o}^o) \quad (7)$$

$$c_{dm} \sim \text{Gauss}(\phi_{z_{dm}^c}^c) \quad (8)$$

$$p_{dm} \sim \text{Gauss}(\phi_{z_{dm}^p}^p) \quad (9)$$

$$a_d \sim \text{Gauss}(\phi_{z_d^a}^a), \quad (10)$$

where the discrete uniform distribution is denoted as $\text{Unif}(\cdot)$, the categorical distribution is denoted as $\text{Cat}(\cdot)$, the Dirichlet distribution is denoted as $\text{Dir}(\cdot)$, the stick-breaking process (SBP) (Sethuraman, 1994) is denoted as $\text{GEM}(\cdot)$, the Gaussian-inverse-Wishart distribution is denoted as $\text{GIW}(\cdot)$, and the multivariate Gaussian distribution is denoted as $\text{Gauss}(\cdot)$. See Murphy (2012) for specific formulas of the above probability distributions. In this paper, variables omitting superscript represent general notation, e.g., $\pi \in \{\pi^a, \pi^p, \pi^o, \pi^c\}$, and variables omitting subscripts represent collective notation, e.g., $F = \{F_1, F_2, \dots, F_D\}$. The number of trials is D . The number of objects on the table is M_d in the d -th trial. The number of words in the sentence is N_d in the d -th trial. The n -th word in the d -th trial is denoted as w_{dn} , which is represented by the bag-of-words (BoW). The model allows sentences containing zero to four words. The model associates the word distributions θ with categories z_{dm} on four sensory-channels, namely, the action a_d , the position p_{dm} of the object on the table, the object feature o_{dm} , and the object color c_{dm} . In this study, we define the action a_d as a static action feature, i.e., proprioceptive and tactile features, when the robot completes

an action. An index of the object of attention selected by the robot from among the multiple objects on the table is denoted as $A_d = m$. The sequence representing the respective sensory-channels associated with each word in the sentence is denoted as F_d , e.g., $F_d = (a, p, c, o)$. The number of categories for each sensory-channel is K . An index of the word distribution is denoted as l . The set of all the word distributions is denoted as $\theta = \{\theta_{l=(F_{dn}, z_{dm}^{F_{dn}})} | F_{dn} \in \{o, c, p, a\}, z_{dm}^{F_{dn}} \in \{1, 2, \dots, K^{F_{dn}}\}\}$. The index of the category of the sensory-channel F_{dn} and the object A_d is denoted as $z_{dA_d}^{F_{dn}}$. Then, the number of word distributions L is the sum of the number of categories of all the sensory-channels, i.e., $L = K^a + K^p + K^o + K^c$. The action category ϕ_k^a , the position category ϕ_k^p , the object category ϕ_k^o , and the color category ϕ_k^c are represented by a Gaussian distribution. The mean vector and the covariance matrix of the Gaussian distribution are denoted as $\phi_k = \{\mu_k, \Sigma_k\}$. We define $\phi_{z_d^a}^a$ as the parameter of the Gaussian distribution that added the object position p_{dA_d} to the element of the mean vector representing the relative coordinates between the hand position and the target position. The position information of the object $A_d = m$ is denoted as p_{dA_d} . Therefore, $\phi_{z_d^a}^a$ is the parameter obtained by converting the target hand position to the absolute coordinate system based on $\phi_{z_d^a}^a$ (the parameter of the action category represented in the relative coordinate system) and p_{dA_d} (the position of the object of attention). The mixture weights of the categories for each sensory-channel are denoted as π^a, π^p, π^o , and π^c . The hyperparameter λ of the uniform distribution, i.e., equation (1), has the mutual exclusivity constraint that determines that each sensory-channel is represented only once in each sentence. The hyperparameter of the mixture weights π is denoted as α . The hyperparameter of the Gaussian-inverse-Wishart distribution is denoted as $\beta = \{\mu_0, \kappa_0, V_0, \nu_0\}$. The hyperparameter of the Dirichlet distribution is denoted as γ . Italic notation (a, p, o, c) represents observation variables, ordinary notation used as a superscript (a, p, o, c) represents sensory-channels.

The robot needs to estimate the number of categories based on experience because the robot cannot have previous knowledge about categories. The proposed method can learn an appropriate number of categories, depending on the collected data, by using a nonparametric Bayesian approach. Specifically, this method uses the SBP, a method based on the Dirichlet process. Therefore, this method can consider theoretically infinite numbers K^a, K^p, K^o , and K^c . In this paper, we approximate the values of parameters representing the number of categories K^a, K^p, K^o , and K^c by assigning sufficiently large values, i.e., a weak-limit approximation (Fox et al., 2011).

3.3. Learning Algorithm

This model estimates parameters representing multiple categories, word distribution, the relationships between the word and the sensory-channel as input for the object features, positions, colors, robot actions, and the sentences spoken by a tutor. The model parameters and latent variables of the proposed method are estimated from the following joint posterior distribution by Gibbs sampling:

$$\Theta, Z \sim p(\Theta, Z | X, H), \quad (11)$$

where the set of model parameters is denoted as $\Theta = \{\pi, \phi, \theta\}$, the set of latent variables is denoted as $Z = \{z, F\}$, the set of

observation variables is denoted as $X = \{a, p, o, c, w, A\}$, and the set of hyperparameters of the model is denoted as $H = \{\{\alpha\}, \{\beta\}, \lambda, \gamma\}$.

The learning algorithm is obtained by repeatedly sampling the conditional posterior distributions for each parameter. The Dirichlet and GIW distributions are conjugate prior distributions for the categorical and Gaussian distributions, respectively (Murphy, 2012). Therefore, the conditional posterior distributions can be determined analytically. **Algorithm 1** shows the pseudo-code for the learning procedure. The initial values of the model parameters can be set arbitrarily in accordance with a condition. The following is the conditional posterior distribution of each element used for performing Gibbs sampling.

A parameter π^o of categorical distribution representing the mixture weight of an object category is sampled as follows:

$$\begin{aligned} \pi^o &\sim p(\pi^o | z^o, \alpha^o) \propto \prod_{d=1}^D \prod_{m=1}^{M_d} \text{Cat}(z_{dm}^o | \pi^o) \text{Dir}(\pi^o | \alpha^o) \\ &\propto \text{Dir}(\pi^o | z^o, \alpha^o), \end{aligned} \quad (12)$$

where z^o denotes the set of all the latent variables of an object category. A parameter π^c of categorical distribution representing the mixture weight of the color category is sampled as follows:

$$\begin{aligned} \pi^c &\sim p(\pi^c | z^c, \alpha^c) \propto \prod_{d=1}^D \prod_{m=1}^{M_d} \text{Cat}(z_{dm}^c | \pi^c) \text{Dir}(\pi^c | \alpha^c) \\ &\propto \text{Dir}(\pi^c | z^c, \alpha^c), \end{aligned} \quad (13)$$

where z^c denotes a set of all the latent variables of the color category. A parameter π^p of the categorical distribution representing the mixture weight of the position category is sampled as follows:

$$\begin{aligned} \pi^p &\sim p(\pi^p | z^p, \alpha^p) \propto \prod_{d=1}^D \prod_{m=1}^{M_d} \text{Cat}(z_{dm}^p | \pi^p) \text{Dir}(\pi^p | \alpha^p) \\ &\propto \text{Dir}(\pi^p | z^p, \alpha^p), \end{aligned} \quad (14)$$

where z^p denotes the set of all the latent variables of the position category. A parameter π^a of the categorical distribution representing the mixture weight of the action category is sampled as follows:

$$\pi^a \sim p(\pi^a | z^a, \alpha^a) \propto \prod_{d=1}^D \text{Cat}(z_d^a | \pi^a) \text{Dir}(\pi^a | \alpha^a) \propto \text{Dir}(\pi^a | z^a, \alpha^a), \quad (15)$$

where z^a denotes a set of all the latent variables of the action category. A parameter ϕ_k^o of the Gaussian distribution of the object category is sampled for each $k \in \{1, 2, \dots, K^o\}$ as follows:

$$\begin{aligned} \phi_k^o &\sim p(\phi_k^o | z^o, o, \beta^o) \propto \prod_{d=1}^D \prod_{m=1}^{M_d} \text{Gauss}(o_{dm} | \phi_k^o) \text{GIW}(\phi_k^o | \beta^o) \\ &\propto \text{GIW}(\phi_k^o | o_k, \beta^o), \end{aligned} \quad (16)$$

where o_k denotes a set of all the object features of the object category $z_{dm}^o = k$ in $m \in \{1, 2, \dots, M_d\}$ and $d \in \{1, 2, \dots, D\}$.

Algorithm 1 | Learning algorithm based on Gibbs sampling.

```

1: procedure Gibbs_Sampling ( $a, p, o, c, w, A$ )
2:   Setting of hyperparameters  $\{\alpha\}, \{\beta\}, \lambda, \gamma$ 
3:   Initialization of parameters and latent variables  $\{\pi\}, \{\phi\}, \theta, \{z\}, F$ 
4:   for  $j = 1$  to iteration_number do
5:      $\pi^o \sim \text{Dir}(\pi^o | z^o, \alpha^o)$  // equation (12)
6:      $\pi^c \sim \text{Dir}(\pi^c | z^c, \alpha^c)$  // equation (13)
7:      $\pi^p \sim \text{Dir}(\pi^p | z^p, \alpha^p)$  // equation (14)
8:      $\pi^a \sim \text{Dir}(\pi^a | z^a, \alpha^a)$  // equation (15)
9:     for  $k = 1$  to  $K^o$  do
10:       $\phi_k^o \sim \text{GIW}(\phi_k^o | o_k, \beta^o)$  // equation (16)
11:     end for
12:     for  $k = 1$  to  $K^c$  do
13:       $\phi_k^c \sim \text{GIW}(\phi_k^c | c_k, \beta^c)$  // equation (17)
14:     end for
15:     for  $k = 1$  to  $K^p$  do
16:       $\phi_k^p \sim \text{GIW}(\phi_k^p | p_k, \beta^p)$  // equation (18)
17:     end for
18:     for  $k = 1$  to  $K^a$  do
19:       $\phi_k^a \sim \text{GIW}(\phi_k^a | a_k, \beta^a)$  // equation (19)
20:     end for
21:     for  $l = (F_{dn}, z_{dn}^{F_{dn}})$  in  $\{(F_{dn}, z_{dn}^{F_{dn}}) | F_{dn} \in \{o, c, p, a\},$ 
22:        $z_{dn}^{F_{dn}} \in \{1, 2, \dots, K^{F_{dn}}\}\}$  do
23:        $\theta_l \sim \text{Dir}(\theta_l | w_l, \gamma)$  // equation (20)
24:     end for
25:     for  $d = 1$  to  $D$  do
26:       for  $m = 1$  to  $M_d$  do
27:          $z_{dm}^o \sim \prod_{n=1}^{N_d} \text{Cat}\left(w_{dn} | \theta_{l=(F_{dn}, z_{dn}^{F_{dn}})}\right) \text{Gauss}\left(o_{dm} | \phi_{z_{dm}^o}^o\right)$ 
28:            $\text{Cat}(z_{dm}^o | \pi^o)$  // equation (21)
29:          $z_{dm}^c \sim \prod_{n=1}^{N_d} \text{Cat}\left(w_{dn} | \theta_{l=(F_{dn}, z_{dn}^{F_{dn}})}\right) \text{Gauss}\left(c_{dm} | \phi_{z_{dm}^c}^c\right)$ 
30:            $\text{Cat}(z_{dm}^c | \pi^c)$  // equation (22)
31:          $z_{dm}^p \sim \prod_{n=1}^{N_d} \text{Cat}\left(w_{dn} | \theta_{l=(F_{dn}, z_{dn}^{F_{dn}})}\right) \text{Gauss}\left(p_{dm} | \phi_{z_{dm}^p}^p\right)$ 
32:            $\text{Cat}(z_{dm}^p | \pi^p)$  // equation (23)
33:       end for
34:        $z_d^a \sim \prod_{n=1}^{N_d} \text{Cat}\left(w_{dn} | \theta_{l=(F_{dn}, z_{dn}^{F_{dn}})}\right) \text{Gauss}\left(a_d | \phi_{z_d^a}^a\right)$ 
35:          $\text{Cat}(z_d^a | \pi^a)$  // equation (24)
36:        $F_d \sim \prod_{n=1}^{N_d} \text{Cat}\left(w_{dn} | \theta_{l=(F_{dn}, z_{dn}^{F_{dn}})}\right) \text{Unif}(F_d | \lambda)$  // equation (25)
37:     end for
38:   return  $\{\pi\}, \{\phi\}, \theta, \{z\}, F$ 
39: end procedure

```

A parameter ϕ_k^c of the Gaussian distribution of the color category is sampled for each $k \in \{1, 2, \dots, K^c\}$ as follows:

$$\begin{aligned} \phi_k^c &\sim p(\phi_k^c | z^c, c, \beta^c) \propto \prod_{d=1}^D \prod_{m=1}^{M_d} \text{Gauss}(c_{dm} | \phi_k^c) \text{GIW}(\phi_k^c | \beta^c) \\ &\propto \text{GIW}(\phi_k^c | c_k, \beta^c), \end{aligned} \quad (17)$$

where c_k denotes the set of all the color features of the color category $z_{dm}^c = k$ in $m \in \{1, 2, \dots, M_d\}$ and $d \in \{1, 2, \dots, D\}$. A parameter ϕ_k^p of the Gaussian distribution of the position category is sampled for each $k \in \{1, 2, \dots, K^p\}$ as follows:

$$\phi_k^p \sim p(\phi_k^p | z^p, p, \beta^p) \propto \prod_{d=1}^D \prod_{m=1}^{M_d} \text{Gauss}(p_{dm} | \phi_k^p) \text{GIW}(\phi_k^p | \beta^p) \\ \propto \text{GIW}(\phi_k^p | p_k, \beta^p), \quad (18)$$

where p_k denotes the set of all the position information of the position category $z_{dm}^p = k$ in $m \in \{1, 2, \dots, M_d\}$ and $d \in \{1, 2, \dots, D\}$. A parameter ϕ_k^a of the Gaussian distribution of the action category is sampled for each $k \in \{1, 2, \dots, K^a\}$ as follows:

$$\phi_k^a \sim p(\phi_k^a | z^a, a, p, A, \beta^a) \propto \prod_{d=1}^D \text{Gauss}(a'_d | \phi_k^a) \text{GIW}(\phi_k^a | \beta^a) \\ \propto \text{GIW}(\phi_k^a | a'_k, \beta^a), \quad (19)$$

where a denotes the set of all the action information, p denotes the set of all the position information, and A denotes the set of all the attention information. The element representing the relative coordinates of the hand of a'_d is calculated by the element representing the absolute coordinates of the hand of a , the object positions p , and the attention information A . The set of all the action information of the action category $z_d^a = k$ in $d \in \{1, 2, \dots, D\}$ is denoted as a'_d . A parameter θ_l of the word probability distribution is sampled for each $l \in \{(F_{dn}, z_{dm}^{F_{dn}}) | F_{dn} \in \{o, c, p, a\}, z_{dm}^{F_{dn}} \in \{1, 2, \dots, K^{F_{dn}}\}\}$ as follows:

$$\theta_l \sim p(\theta_l | w, z^o, z^c, z^p, z^a, F, A, \gamma) \\ \propto \prod_{d=1}^D \prod_{n=1}^{N_d} \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dm}^{F_{dn}})}) \text{Dir}(\theta_l | \gamma) \\ \propto \text{Dir}(\theta_l | w_l, \gamma) \quad (20)$$

where w denotes the set of all the words, F denotes the set of frames of all the sentences, and w_l denotes the set of all the words of the word category $l = (F_{dn}, z_{dm}^{F_{dn}})$ in $n \in \{1, 2, \dots, N_d\}$ and $d \in \{1, 2, \dots, D\}$. A latent variable z_{dm}^o of the object category is sampled for each $m \in \{1, 2, \dots, M_d\}$ and $d \in \{1, 2, \dots, D\}$ as follows:

$$z_{dm}^o \sim p(z_{dm}^o | w_d, z_d^c, z_d^p, z_d^a, z_{-dm}^o, \theta, F_d, A_d, o_{dm}, \phi^o, \pi^o) \\ \propto \prod_{n=1}^{N_d} \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dm}^{F_{dn}})}) \text{Gauss}(o_{dm} | \phi_{z_{dm}^o}^o) \text{Cat}(z_{dm}^o | \pi^o), \quad (21)$$

where w_d is a sequence of words in the d -th trial and z_{-dm}^o is the set of indicators of the object categories without z_{dm}^o in the d -th trial. A latent variable z_{dm}^c of the color category is sampled for each $m \in \{1, 2, \dots, M_d\}$ and $d \in \{1, 2, \dots, D\}$ as follows:

$$z_{dm}^c \sim p(z_{dm}^c | w_d, z_d^o, z_d^p, z_d^a, z_{-dm}^c, \theta, F_d, A_d, c_{dm}, \phi^c, \pi^c) \\ \propto \prod_{n=1}^{N_d} \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dm}^{F_{dn}})}) \text{Gauss}(c_{dm} | \phi_{z_{dm}^c}^c) \text{Cat}(z_{dm}^c | \pi^c), \quad (22)$$

where z_{-dm}^c is the set of indicators of the object categories without z_{dm}^c in the d -th trial. A latent variable z_{dm}^p of the position category is sampled for each $m \in \{1, 2, \dots, M_d\}$ and $d \in \{1, 2, \dots, D\}$ as follows:

$$z_{dm}^p \sim p(z_{dm}^p | w_d, z_d^o, z_d^c, z_d^a, z_{-dm}^p, \theta, F_d, A_d, p_{dm}, \phi^p, \pi^p) \\ \propto \prod_{n=1}^{N_d} \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dm}^{F_{dn}})}) \text{Gauss}(p_{dm} | \phi_{z_{dm}^p}^p) \text{Cat}(z_{dm}^p | \pi^p), \quad (23)$$

where z_{-dm}^p is the set of indicators of the object categories without z_{dm}^p in the d -th trial. A latent variable z_d^a of the action category is sampled for each $d \in \{1, 2, \dots, D\}$ as follows:

$$z_d^a \sim p(z_d^a | w_d, z_d^o, z_d^c, z_d^p, \theta, F_d, A_d, a_d, p_d, \phi^a, \pi^a) \\ \propto \prod_{n=1}^{N_d} \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dm}^{F_{dn}})}) \text{Gauss}(a_d | \phi_{z_d^a}^a) \text{Cat}(z_d^a | \pi^a), \quad (24)$$

where p_d is the set of position data in the d -th trial. A latent variable F_d representing the sensory-channels of words in a sentence is sampled for each $d \in \{1, 2, \dots, D\}$ as follows:

$$F_d \sim p(F_d | w, z^o, z^c, z^p, z^a, \theta, A, \lambda) \\ \propto \prod_{n=1}^{N_d} \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dm}^{F_{dn}})}) \text{Unif}(F_d | \lambda). \quad (25)$$

3.4. Action Generation and Attention Selection

In this section, we describe the approach that selects an action and an object of attention from the human spoken sentence. A robot capable of learning word meanings accurately is considered to be able to understand human instruction more accurately. In an action generation task, the robot performs an action a_d based on word meanings and multiple categories Θ from observed information w_d , o_d , c_d , and p_d . In this case, the robot can use the set of model parameters Θ learned by using Gibbs sampling in the CSL task. In the action generation task, we maximize the following equation:

$$\arg\max_{a_d} p(a_d | w_d, o_d, c_d, p_d, \theta, \{\phi\}, \{\pi\}, \lambda) \\ = \arg\max_{a_d} \sum_{A_d} \sum_{z_d^a} p(a_d | \phi^a, z_d^a, p_d, A_d) \\ \times p(A_d, z_d^a | w_d, o_d, c_d, p_d, \theta, \{\phi\}, \{\pi\}, \lambda). \quad (26)$$

In practice, this maximization problem is separated into two approximation processes, because it is difficult to maximize equation (26) directly.

- (1) The first process is the maximization of the attention A_d and the index of the action category z_d^a

$$A_d^*, z_d^{a*} = \arg\max_{A_d, z_d^a} p(A_d, z_d^a | w_d, o_d, c_d, p_d, \theta, \{\phi\}, \{\pi\}, \lambda). \quad (27)$$

The probability distribution of equation (27) is represented by the following equation:

$$\begin{aligned}
 & p(A_d, z_d^a | w_d, o_d, c_d, p_d, \theta, \{\phi\}, \{\pi\}, \lambda) \\
 & \propto p(A_d = m) p(z_d^a | \pi^a) \prod_{M_d} \sum_{z_{dm}^o} \sum_{z_{dm}^c} \sum_{z_{dm}^p} \\
 & \text{Gauss}(o_{dm} | \phi_{z_{dm}^o}^o) \text{Cat}(z_{dm}^o | \pi^o) \\
 & \text{Gauss}(c_{dm} | \phi_{z_{dm}^c}^c) \text{Cat}(z_{dm}^c | \pi^c) \\
 & \text{Gauss}(p_{dm} | \phi_{z_{dm}^p}^p) \text{Cat}(z_{dm}^p | \pi^p) \\
 & \left[\sum_{F_d} \text{Unif}(F_d | \lambda) \prod_{N_d} \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dA_d}^{F_{dn}})}) \right]. \quad (28)
 \end{aligned}$$

Then, we assumed $p(A_d = m) = 1/M_d$ as equal probability for the number of objects.

- (2) The second process is the maximization of the action a_d using A_d^* and z_d^{a*}

$$\begin{aligned}
 a_d^* &= \underset{a_d}{\operatorname{argmax}} p(a_d | \phi^a, z_d^{a*}, p_d, A_d^*) \\
 &= \underset{a_d}{\operatorname{argmax}} \text{Gauss}(a_d | \phi_{z_d^{a*}}^{a'}) \\
 &= \mu_{z_d^{a*}}^{a'}, \quad (29)
 \end{aligned}$$

where the mean vector of the Gaussian distribution of the action category z_d^{a*} is denoted as $\mu_{z_d^{a*}}^{a'}$.

3.5. Description of the Current Situation and Self-Action by the Robot

In this section, we describe the approach followed by the description task representing the current situation and the self-action of the robot. We consider a robot capable of learning word meanings accurately to be able to describe the current situation and self-action more accurately. In the action description task, the robot utters a sentence w_d regarding a self-action a_d and observed information o_d , c_d , and p_d based on word meanings and multiple categories Θ . In this case, the robot can use the set of model parameters Θ learned by using Gibbs sampling in the CSL task. In the action description task, we maximize the following equation:

$$\begin{aligned}
 & \underset{w_d}{\operatorname{argmax}} p(w_d | a_d, o_d, c_d, p_d, \theta, \{\phi\}, \{\pi\}, F_d, A_d) \\
 & \propto \underset{w_d}{\operatorname{argmax}} \sum_{z_d^a} \sum_{z_{dA_d}^o} \sum_{z_{dA_d}^c} \sum_{z_{dA_d}^p} \\
 & \text{Gauss}(a_d | \phi_{z_d^a}^{a'}) \text{Cat}(z_d^a | \pi^a), \\
 & \text{Gauss}(o_{dA_d} | \phi_{z_{dA_d}^o}^o) \text{Cat}(z_{dA_d}^o | \pi^o) \\
 & \text{Gauss}(c_{dA_d} | \phi_{z_{dA_d}^c}^c) \text{Cat}(z_{dA_d}^c | \pi^c) \\
 & \text{Gauss}(p_{dA_d} | \phi_{z_{dA_d}^p}^p) \text{Cat}(z_{dA_d}^p | \pi^p) \\
 & \prod_{N_d} \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dA_d}^{F_{dn}})}). \quad (30)
 \end{aligned}$$

If the frame of the sentence is decided, e.g., $F_d = (a, p, c, o)$, equation (30) is represented as the following:

$$\begin{aligned}
 \text{Equation (30)} &= \prod_{N_d} \underset{w_{dn}}{\operatorname{argmax}} \sum_{z_{dA_d}^{F_{dn}}} \text{Gauss}(x_{dA_d}^{F_{dn}} | \phi_{z_{dA_d}^{F_{dn}}}^{F_{dn}}) \\
 &\times \text{Cat}(z_{dA_d}^{F_{dn}} | \pi^{F_{dn}}) \text{Cat}(w_{dn} | \theta_{l=(F_{dn}, z_{dA_d}^{F_{dn}})}), \quad (31)
 \end{aligned}$$

where $x_{dA_d}^{F_{dn}}$ denotes data of the sensory-channel F_{dn} in the object number A_d , i.e., a_d , p_{dA_d} , c_{dA_d} , or o_{dA_d} . Therefore, equation (30) can be divided into the equations of finding a maximum value for each word.

4. EXPERIMENT I: SIMULATION ENVIRONMENT

We performed the experiments described in this section using the iCub simulator (Tikhonoff et al., 2008). In Section 4.1, we describe the difference in the conditions of the methods that are used for comparison purposes. In Section 4.2, we describe the CSL experiment. In Section 4.3, we describe the experiment involving the action generation task. In Section 4.4, we describe the experiment relating to the action description task.

4.1. Comparison Methods

We evaluated our proposed method by comparing its performance with that of two other methods.

- (A) The proposed method.

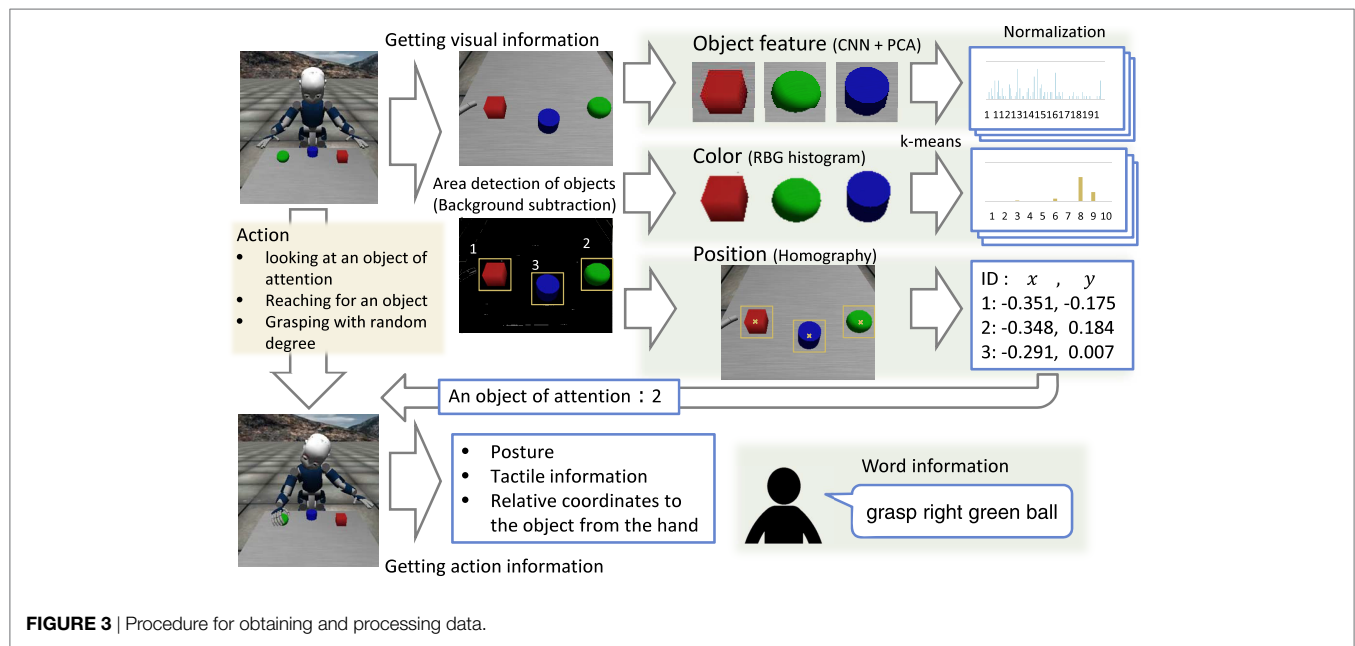
This method has a mutual exclusivity constraint between the word and the sensory-channel (MEC-I and II), determining that each sensory-channel occurs only once in each sentence. For example, if the number of words in a sentence is $N_d = 4$, F_d can become a sequence such as (a, c, p, o), (a, p, c, o), or (p, c, o, a). Possible values of F_d are constrained by λ as a permutation of four sensory-channels. The number of permutations is ${}_4P_{N_d} = 4!/(4 - N_d)!$.

- (B) The proposed method without the mutual exclusivity constraint (w/o MEC-II).

This method does not have the mutual exclusivity constraint (MEC-II). This means that several words in a sentence may relate to the same sensory-channel. For example, if the number of words in a sentence is $N_d = 4$, F_d can become a sequence such as (a, o, c, o), (a, p, p, o), or (o, o, o, o) in addition to the above example of (A). Possible values of F_d are constrained by λ as a repeated permutation of four sensory-channels. The number of repeated permutations is ${}_4\Pi_{N_d} = 4^{N_d}$. In this case, the robot needs to consider additional pairs of relationships between the sensory-channel and word compared to method (A).

- (C) The multilayered multimodal latent Dirichlet allocation (mMLDA) (Attamimi et al., 2016).

This method is based on mMLDA. In this research, this method was modified from the actual mMLDA to apply to our task and the proposed method. In particular, the emission probability for each sensory-channel is changed from



a categorical distribution to a Gaussian distribution. This means the multimodal categorization methods are based on a Gaussian distribution for each sensory-channel and a categorical distribution for word information. This method relates all observed words in a situation to all observed sensory-channel information in the situation. This method neither has the mutual exclusivity constraint (MEC-I and II) nor does it select the sensory-channel by words, i.e., F_d is not estimated.

4.2. Cross-Situational Learning

4.2.1. Experimental Procedure and Conditions

We conducted an experiment to learn the categories for each sensory-channel and the words associated with each category. **Figure 3** shows the procedure for obtaining and processing data. We describe the experimental procedure for CSL as follows:

1. The robot takes the initial position and posture. Some objects are placed on the table.
2. The robot acquires a visual image of the table. Subsequently, the robot detects object areas by using background subtraction. The detected object areas are cut out as object images of 64×64 pixels. The robot obtains the number of objects on the table.
3. The robot extracts object features, color features, and object positions. We used the deep learning framework Caffe (Jia et al., 2014) for convolutional neural networks (CNNs) (Krizhevsky et al., 2012) as an object feature extractor. We used a pre-trained CNN, i.e., CaffeNet trained by using ImageNet Large Scale Visual Recognition Challenge 2012 as the dataset. The object features are obtained from the fully connected FC6 layer (4096-dimensions) in CaffeNet. After that, the object features are reduced by principal component analysis (PCA). In terms of color features, the RGB histogram is vector quantized by k-means and normalized. The position data are converted into the world coordinate by homography. The position data are two-dimensional to represent the plane of the table.
4. The robot performs an action including a little randomness to an object of attention. The difference and uncertainty in the robot's action are represented by this randomness. First, the robot moves its eye-gaze to an object of attention. The object is selected randomly. Next, the robot moves its right hand to the coordinates of the target object by using inverse kinematics. A little random noise is added to a target position of the end-effector of the right hand. In many cases, the robot moves its right hand after looking at the object. The robot rarely refrains from moving its hands, looks at the object, which means the action of "look-at." When the hand approaches the position of the target object, the robot bends its fingers. The rate at which it bends its fingers is selected randomly. The five fingers move in synchronization. After the action is completed, the robot acquires the data relating to this action, including data relating to the posture, tactile data, and the relative coordinates of the object from its right hand. The action data are 38-dimensional and include the position of the right hand relative to the object (3-dim.), the rate at which the finger bends (1-dim.), the joint angles of the head, right_arm, and torso (6, 16, 3-dim.), and tactile information of the right hand (9-dim.). The action data is normalized to [0,1] for each dimension.
5. When the robot completes an action, the human tutor speaks a sentence about the object of attention and the action of the robot. The sentence contains the word related to each sensory-channel once, e.g., "touch left red box." In this task, the number of words in the sentence is indicated by a number ranging from zero to four. Zero means that the tutor did not speak a sentence.

The above process is carried out many times in different situations. The robot learns multiple categories and word meanings by using multimodal data observed in many trials.

The number of trials was $D = 20$ and 40 for CSL. The number of objects M_d on the table for each trial was a number from

one to three. The number of words N_d in the sentence was a number from zero to four. We assume that a word related to each sensory-channel is spoken only once in each sentence. The word order in the sentences was changed. This experiment used 14 kinds of words: “reach,” “touch,” “grasp,” “look-at,” “front,” “left,” “right,” “far,” “green,” “red,” “blue,” “box,” “cup,” and “ball.” The upper limit number of the categories for each sensory-channel was $K = 10$, i.e., the number of word distributions was $L = 40$. The number of iterative cycles used for Gibbs sampling was 200. The hyperparameters were $\alpha = 1.0$, $\gamma = 0.1$, $m_0 = \mathbf{O}_{x_{dim}}$, $\kappa_0 = 0.001$, $V_0 = \text{diag}(0.01, 0.01)$, and $v_0 = x_{dim} + 2$, where the number of dimensions for each sensory-channel x is denoted as x_{dim} and the zero vector in x_{dim} dimensions is denoted as $\mathbf{O}_{x_{dim}}$. PCA is used to reduce the object features to 30 dimensions. The color features are quantized to 10 dimensions by k-means.

We describe the criteria of words uttered for action category as follows: “reach” corresponds to the robot extending its right hand toward an object and the robot’s finger does not make contact with an object; “touch” corresponds to the robot touching an object and its finger is relatively opened; “grasp” corresponds to the robot’s hand holding firmly an object; “look-at” corresponds to the robot not moving its right hand and it focuses on an object of attention only. Based on these criteria, the tutor determines an action word. In particular, “reach” and “touch” are similar; the only difference is whether the hand touches the object or not.

We evaluate the estimation accuracy of the learning results by using uncertain teaching sentences. Each sentence contains four words or fewer in different order. We compare the accuracy of three methods by reducing the word information. In addition, the number of learning trials is changed. We compared the accuracy by changing the number of trials. We evaluated the methods according to the following metrics.

- Adjusted Rand index (ARI)

We compare the matching rate between the estimated latent variables z for each sensory-channel and the true categorization results. The evaluation of this experiment uses the ARI (Hubert and Arabie, 1985), which is a measure of the degree of similarity between two clustering results.

- Estimation accuracy rate of F_d (EAR)

The evaluation of the estimation results of the sensory-channels corresponding to the words are determined as follows:

$$\text{EAR} = 1 - \frac{\text{The number of estimation errors}}{\text{The number of all of uttered words}}. \quad (32)$$

4.2.2. Learning Results and Evaluation

The learning results obtained by using the proposed method are presented here. Forty trials were used. In this case, the number of words was four in all utterance sentences. **Figure 4A** shows the word probability distributions θ . Higher probability values are represented by darker shades. If the relationship between the word and sensory-channel can be estimated correctly, the ranges within thick-bordered boxes show higher probabilities. For example, the action categories show higher probabilities for words of action (“touch,” “look-at,” “reach,” and “grasp”). The categories of the other sensory-channels are also the same. In the position and color

categories, the estimated number of categories was equal to the number of types of words representing the sensory-channel. In the action category, the words “touch,” “reach,” and “grasp” were associated across several categories. In addition, these words were confused with each other. We considered actions representing these words to be ambiguous and similar. On the other hand, we considered the reason why these actions were divided into several categories to be a change in posture information depending on the position of the target object. **Figure 4B** shows the learning result for the position category ϕ^p . For example, the position category p1 is associated with the word “front” (see **Figures 4A,B**). **Figures 4C,D** show examples of the categorization results for the object and color categories. The object categorization result was not perfect. We considered the robot to find it difficult to clearly distinguish objects of different shapes because the 3D-models of the objects had simple shapes. The color categorization result was perfect. In this case, F_d was correctly estimated in all of the trials. The results demonstrate that the proposed method was able to accurately associate each word with its respective sensory-channel.

We performed the learning scenarios 10 times for each method. **Tables 1A,B** show the evaluation values of the experimental results for 20 and 40 trials. The rate of omitted words (ROW), which is expressed as a percentage, represents the uncertainty of teaching sentences. For example, the total number of words is 80 when ROW is 0%, 64 words for 20%, 48 words for 40%, and 32 words for 60% in 20 trials. Also, the total number of words is 160 for a ROW value of 0% and 96 words for 40% in 40 trials. ARI_a, ARI_p, ARI_o, and ARI_c are the ARI values of the action, position, object, and color category, respectively. The EAR values of mMLDA were not calculated because this method does not have F_d . If the ROW value is 100 (no word), the three methods will be equivalent as ALL, i.e., the GMM for each sensory-channel. We described the ARI values of ALL as reference values because ALL is not CSL. The EAR value obtained for the proposed method was higher than that obtained for the other methods. When the ROW decreased, i.e., the word information increased, the evaluation values tended to increase. Particularly, the result for the position category was favorably affected by the increase in word information for categorization. In addition, when the number of trials increased, the evaluation values tended to increase. This result suggests that the robot is able to learn the word meanings more effectively by accumulating more experience even in more complicated and uncertain situations. When the number of words was small (i.e., the ROW value is 40 or 60%), the difference between the EAR values of methods (A) and (B) was small (approximately 0.02) in 20 trials. However, when the number of words was large, the difference between the EAR values of methods (A) and (B) increased, and the EAR value of the method (A) was larger than that of (B). As a result, when the number of words was small, e.g., sentences including one or two words, there was almost no influence of the presence or absence of the MEC-II because the number of possible values of F_d of the methods (A) and (B) were close. On the other hand, when the number of words was large, e.g., sentences including four words, the MEC-II worked well because the number of possible values of F_d of the method (A) was narrowed properly down.

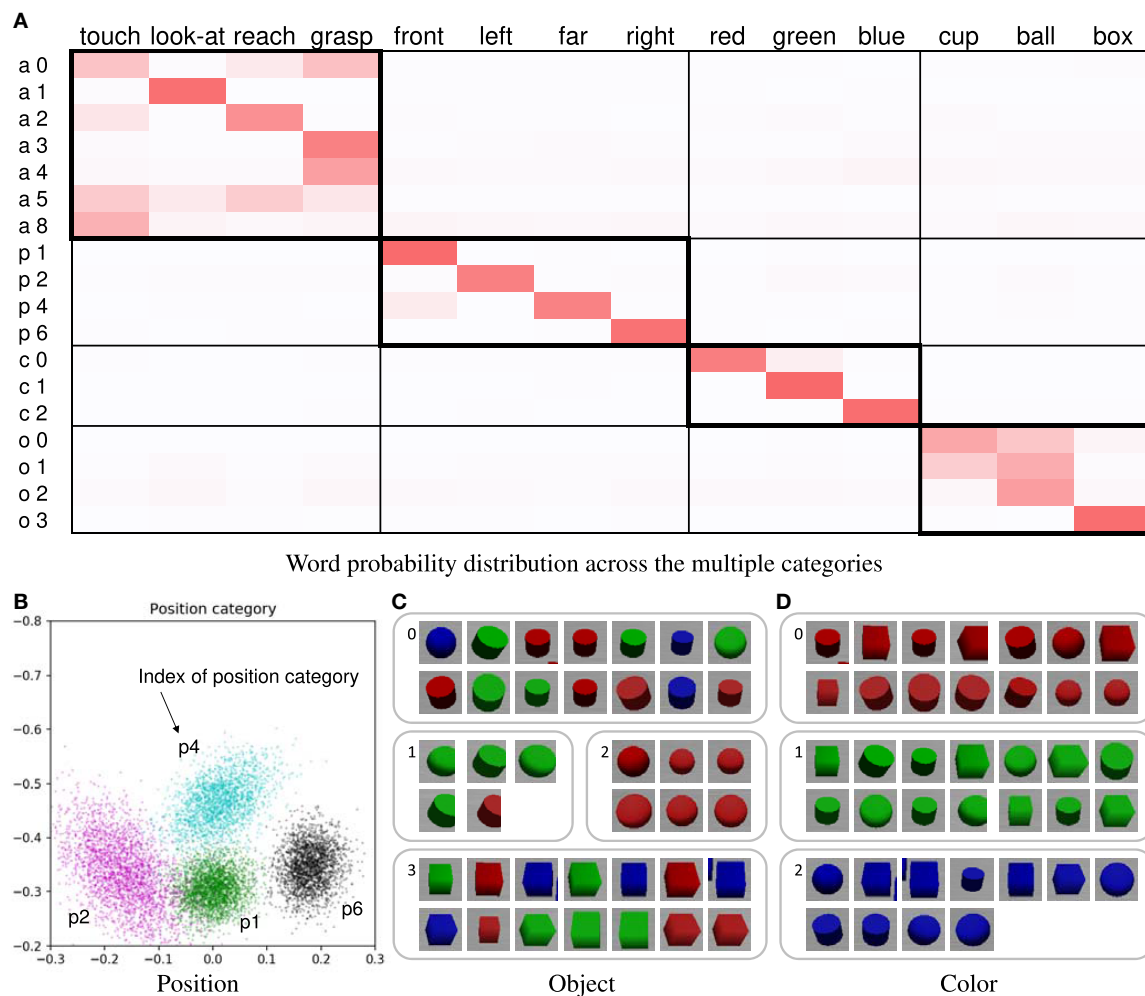


FIGURE 4 | (A) Word probability distribution across the multiple categories; darker shades represent higher probability values. The pair consisting of a letter and a number on the left of the table is the index of the word distribution, which represents the sensory-channel related to the word distribution and the index of the category. Note that category indices are not shown; they are merged and not used because the number of the categories is automatically estimated by the nonparametric Bayesian method. **(B)** Learning result of the position category; for example, the index of position category p1 corresponds to the word “front.” The point group of each color represents each Gaussian distribution of the position category. The crosses in the different colors represent the object positions of the learning data. Each color represents a position category. **(C)** Example of categorization results of object category; **(D)** example of categorization results of color category.

4.3. Action Generation Task

4.3.1. Experimental Procedure and Conditions

In this experiment, the robot generates the action regarding the sentence spoken by the human tutor. The robot uses the learning results of the CSL task in Section 4.2. The robot selects the object of attention from among the objects on the table. In addition, the robot performs the action on the object of attention. In this task, the robot cannot use joint attention. Therefore, the robot needs to overcome both the problems of CSL-I and II. We describe the process of action generation as follows:

1. The robot takes the initial position and posture. Some objects are placed on the table.
2. The tutor speaks a sentence about an action that should be performed by the robot. The robot recognizes the tutor’s spoken sentence.

3. The robot detects the objects on the table. The robot obtains object, color features, and position data by the same process as in Section 4.2.1 (step 3).
4. The robot selects the action category and the object of the attention by using equation (26). The robot calculates the target position by using equation (29).
5. The robot directs its eye-gaze to the object of attention, and the robot performs an action on the object of attention.

The above process is carried out many times on different sentences.

We compare the three methods by quantitative evaluation on the action generation task. We evaluate the accuracy of the selection of the object of attention. In addition, we evaluate the accuracy of an action of the robot based on questionnaire evaluation by participants. The robot generates an action from the tutor’s spoken sentence in a situation. Participants check videos of the action

generated by the robot and select a word representing the robot's action. We calculate the word accuracy rate (WAR) of the words selected by participants and the true words spoken by the tutor. In addition, we calculate the object accuracy rate (OAR) representing the rate at which the robot correctly selected the object instructed by the tutor.

We performed action generation tasks for a total of 12 different test-sentences. The test-sentences included four words representing the four sensory-channels. This placement of objects on the table was not used during the learning trials. In addition, the word order of sentences uttered during the action generation task is different from the word order of sentences uttered during the CSL task. The eight participants checked 36 videos of the robot's actions.

4.3.2. Results

Figure 5 shows three examples of the action generation results of the proposed method. **Figure 5A** shows the result of action generation by the robot in response to the sentence “reach front

blue cup.” **Figure 5B** shows the result of action generation by the robot in response to the sentence “grasp right green ball.” **Figure 5C** shows the result of action generation by the robot in response to the sentence “touch left red box.” **Table 2** shows the results of the quantitative evaluation of the action generation task. The proposed method enabled the robot to accurately select the object. As a result of the proposed method, the object indicated and the object selected by the robot coincided in all sentences. In addition, the proposed method showed the highest values for both WAR and OAR. Therefore, the robot could select an appropriate object and could perform an action even in situations and for sentences not used for CSL.

4.4. Action Description Task

4.4.1. Experimental Procedure and Conditions

In HRI, the ability of the robot to use the acquired word meanings for a description of the current situation is important. In this experiment, the robot performs an action and speaks the sentence

TABLE 1 | Experimental results of the CSL task for 20 and 40 trials.

Method	ROW	ARI_a	ARI_p	ARI_o	ARI_c	EAR_F _d
(A) 20 trials						
Proposed	0	0.300	0.606	0.408	0.782	<u>0.970</u>
w/o MEC-II	0	0.317	0.648	0.338	0.805	<u>0.759</u>
mMLDA	0	0.316	0.428	0.277	0.756	–
Proposed	20	0.290	0.564	0.332	0.762	0.727
w/o MEC-II	20	0.342	0.486	0.436	0.755	0.598
mMLDA	20	0.267	0.494	0.369	0.776	–
Proposed	40	0.324	0.493	0.354	0.780	0.556
w/o MEC-II	40	0.318	0.486	0.347	0.812	0.529
mMLDA	40	0.356	0.479	0.312	0.771	–
Proposed	60	0.282	0.460	0.295	0.783	0.381
w/o MEC-II	60	0.311	0.454	0.326	0.750	0.406
mMLDA	60	0.294	0.487	0.403	0.724	–
ALL	100 (no word)	0.325	0.431	0.346	0.751	–
(B) 40 trials						
Proposed	0	0.375	0.540	0.366	0.870	<u>0.989</u>
w/o MEC-II	0	0.383	0.524	0.333	0.805	0.834
mMLDA	0	0.388	0.594	0.377	0.822	–
Proposed	40	0.368	0.543	0.313	0.835	<u>0.867</u>
w/o MEC-II	40	0.417	0.577	0.320	0.842	0.780
mMLDA	40	0.340	0.600	0.377	0.856	–

Bold and underscore indicate the highest evaluation values, and bold indicates the second highest evaluation values.

TABLE 2 | Results of evaluation values for the action generation using the results of the CSL for 40 trials (ROW is 0%).

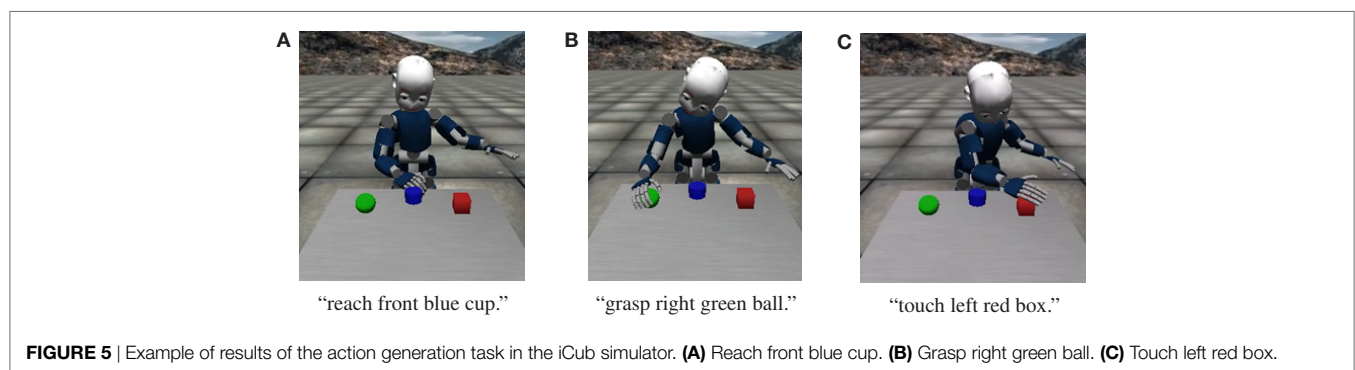
Method	WAR	OAR
Proposed	<u>0.604</u>	<u>1.000</u>
w/o MEC-II	<u>0.510</u>	<u>0.917</u>
mMLDA	0.260	0.667

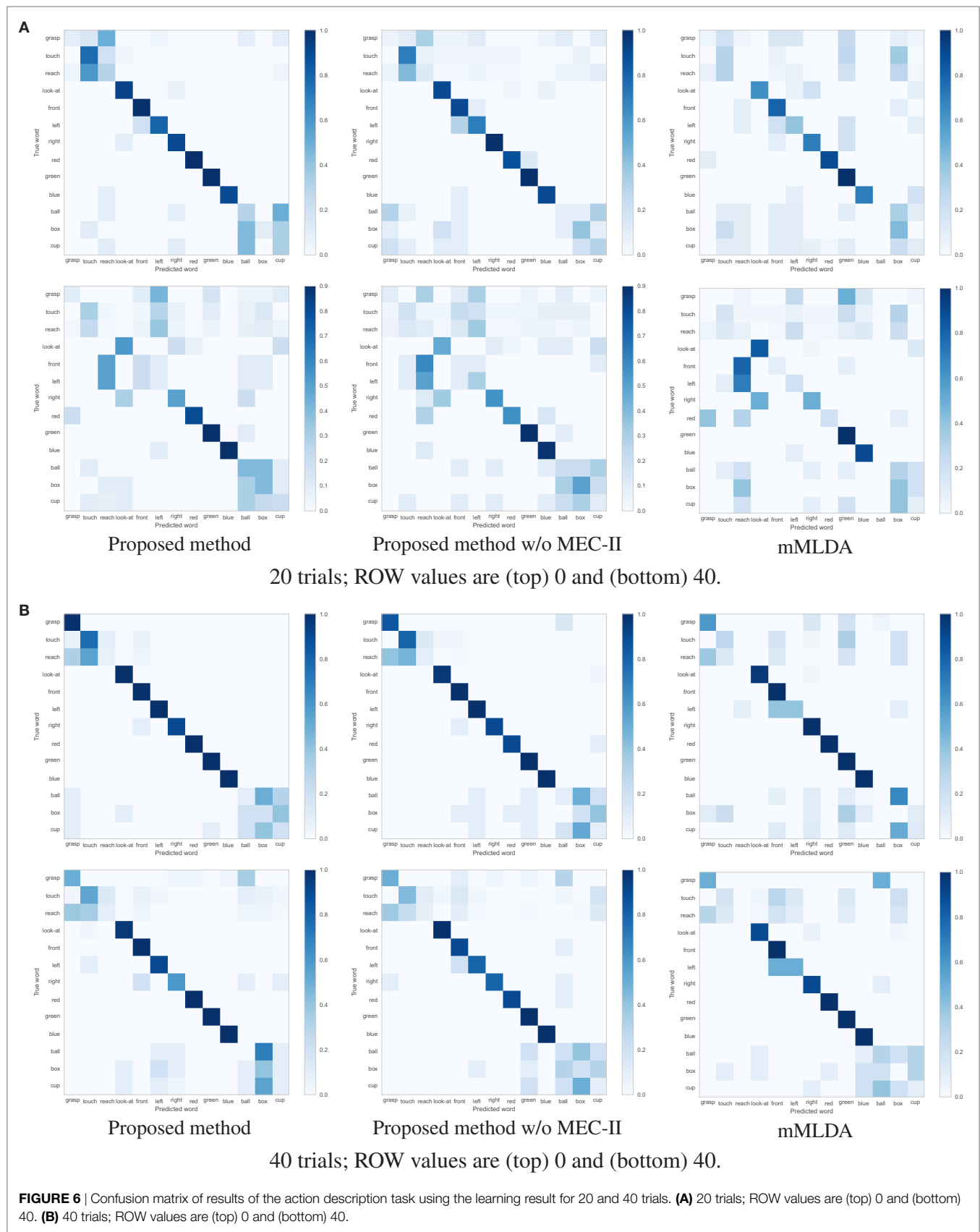
Bold and underscore indicate the highest evaluation values, and bold indicates the second highest evaluation values.

TABLE 3 | Experimental results of action description task for 20 and 40 trials.

Method	Trials	ROW	F1	ACC
Proposed	20	0	<u>0.586</u>	<u>0.660</u>
w/o MEC-II	20	0	<u>0.534</u>	<u>0.613</u>
mMLDA	20	0	0.401	0.469
Proposed	20	40	<u>0.388</u>	<u>0.425</u>
w/o MEC-II	20	40	<u>0.343</u>	<u>0.369</u>
mMLDA	20	40	0.319	0.352
Proposed	40	0	<u>0.663</u>	<u>0.692</u>
w/o MEC-II	40	0	<u>0.642</u>	<u>0.671</u>
mMLDA	40	0	0.474	0.560
Proposed	40	40	<u>0.588</u>	<u>0.623</u>
w/o MEC-II	40	40	<u>0.548</u>	<u>0.606</u>
mMLDA	40	40	0.479	0.569

Bold and underscore indicate the highest evaluation values, and bold indicates the second highest evaluation values.





corresponding to this action. In other words, the robot explains self-action by using a sentence. The robot uses the learning results of the CSL task in Section 4.2. We describe the process of action description as follows:

1. The robot takes the initial position and posture. Some objects are placed on the table.
2. The robot detects the objects on the table. The robot obtains the object, color features, and position data by the same process as in Section 4.2.1 (step 3).
3. The robot selects the object of attention randomly. The robot directs its eye-gaze to the object of attention, and the robot performs an action on the object of attention by using the same process as in Section 4.2.1 (step 4).
4. When the robot completes an action, it utters a sentence about this action.

The above process is carried out many times on different actions. We performed action description tasks for a total of 12 actions. This placement of objects on the table was not used during the learning trials. The robot generates a sentence consisting of four words that include the four sensory-channels. The word order in the sentence is fixed as $F_d = (a, p, c, o)$.

We compare the three methods by quantitative evaluation of the action description task. We evaluate the F1-measure and the accuracy (ACC) between the sentence generated by the robot and the correct sentence decided by the tutor. The evaluation values are calculated by generating the confusion matrix between the predicted words and true words.

4.4.2. Results

Table 3 show the F1-measure and ACC values of the action description task using the learning results under the different conditions. The proposed method showed the highest evaluation values. Figures 6A,B shows the confusion matrices of the results of the action description task using the learning result for 20 and 40 training trials. Overall, the robot confused the words “reach” and “touch” similar to the learning result in Figure 4A. The robot had difficulties in distinguishing between “reach” and “touch.” In other words, this result suggests that these words were learned as synonyms. When the ROW increased, the evaluation values decreased. For the ROW value of 40% obtained for 20 trials, the robot confused words related to the action and position categories. This could be explained by considering that the robot misunderstood the correspondence between the word and the sensory-channel because the word information was insufficient and uncertain during CSL with the ROW value of 40% and 20 trials. On the other hand, an increase in the number of learning trials resulted in an increase in the evaluation values. Even if the robot is exposed to uncertain utterances, the robot can explain self-action more accurately by gaining more experience. As a result, the robot could acquire the ability to explain self-action by CSL based on the proposed method.

5. EXPERIMENT II: REAL iCub ENVIRONMENT

In this section, we describe the experiment that was conducted by using the real iCub robot. The real-world environment involves

more complexity than the simulation environment. We demonstrate that results similar to those of the simulator experiment can be obtained even in a more complicated real environment. We compare three methods, as in Section 4.1. In Section 5.1, we describe the experiment to assess cross-situational learning. In Section 5.2, we describe the experiment relating to the action generation task. In Section 5.3, we describe the experiment relating to the action description task.

5.1. Cross-Situational Learning

5.1.1. Conditions

The experimental procedure is the same as in Section 4.2.1. We use ARI and EAR as evaluation values. Figure 7 shows all of the objects that were used in the real environment. We used 14 different objects including four types (car, cup, ball, and star) and four colors (red, green, blue, and yellow). In the simulation environment, the same type objects had the same shapes. In the real environment, objects of the same type include different shapes. In particular, all the car objects have different shapes, the cup objects have different sizes, and the star objects include one different shape. This experiment used 16 kinds of words: “reach,” “touch,” “grasp,” “look-at,” “front,” “left,” “right,” “far,” “green,” “red,” “blue,” “yellow,” “car,” “cup,” “ball,” and “star.” The number of trials was $D = 25$ and 40 for CSL. The number of objects M_d on the table for each trial was a number ranging from one to three. The number of words N_d in the sentence was a number ranging from zero to four. We assume that a word related to each sensory-channel is spoken only once in each sentence. The word order in the sentences was changed. Object features are reduced to 65 dimensions by PCA. Color features are quantized to 10 dimensions by k-means. The upper limit number of the categories for each sensory-channel was $K = 10$, i.e., the upper limit for the number of word distributions was $L = 40$. The hyperparameters were $\alpha = 1.0$, $\gamma = 0.1$, $m_0 = \mathbf{O}_{x_{dim}}$, $\kappa_0 = 0.001$, $V_0 = \text{diag}(0.01, 0.01)$, and $v_0 = x_{dim} + 2$. The number of iterative cycles used for Gibbs sampling was 200.

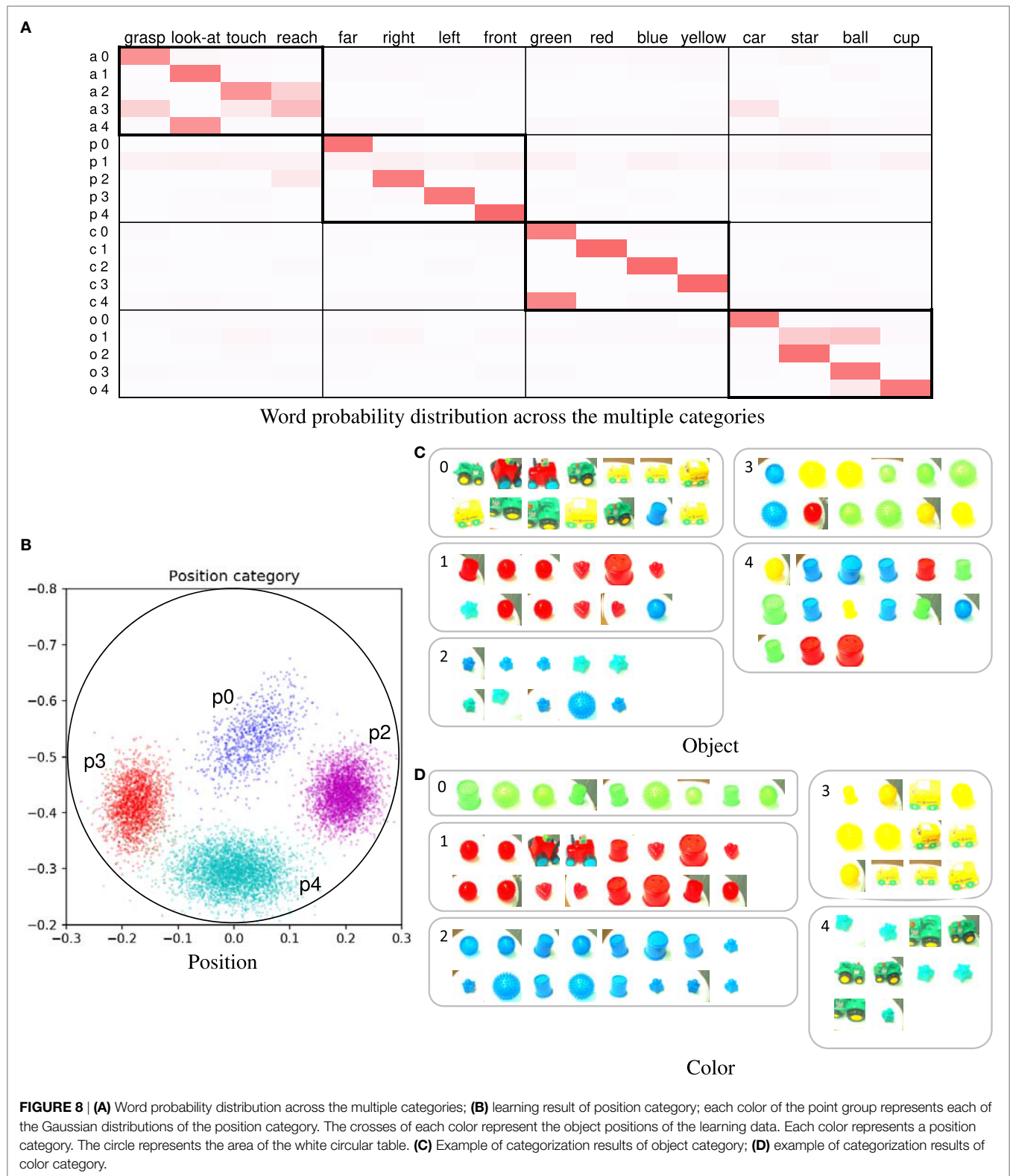


FIGURE 7 | All of the objects used in the real experiments (14 objects including four types and four colors).

5.1.2. Learning Results and Evaluation

The example we describe is the learning result of 25 trials and for a ROW value of 9%. In this case, the number of categories was set to $K = 5$. **Figure 8A** shows the word distributions θ . In the action category, the robot confused the words “reach” and “touch”

as is the case with the simulator experiment. **Figure 8B** shows the learning result of the position category on the table. **Figure 8C** shows categorization results of objects. Although the object categorization contained a few mistakes, the results were mostly correct. **Figure 8D** shows the categorization results obtained for the



color categorization, which was successful. Interestingly, two categories corresponding to the word “green” were created because the robot distinguished between bright green and dark green. In addition, the robot was able to learn that both of these categories related to the word “green.”

Table 4 shows the evaluation values of the experimental results for 25 and 40 trials. There was not much difference in ARI values between the methods and between different conditions of ROW values. The EAR values of the proposed method were higher than those of the other methods. An increase in the number of trials led to an increase in the evaluation values, similar to the simulation results.

5.2. Action Generation Task

5.2.1. Conditions

In this experiment, the robot generates the action corresponding to the sentence spoken by the human tutor. The robot uses the learning results of the CSL task in Section 5.1. The experimental procedure is the same as in Section 4.3.1. We evaluated accuracy of object selection (the OAR values) using the CSL results for 25 trials. We performed the action generation task for

TABLE 4 | Experimental results of the CSL task for 25 and 40 trials.

Method	ROW	ARI_a	ARI_p	ARI_o	ARI_c	EAR_F _d
(A) 25 trials						
Proposed	0	0.239	0.932	0.201	0.720	<u>0.866</u>
w/o MEC-II	0	0.299	0.971	0.207	0.717	0.723
mMLDA	0	0.255	0.959	0.226	0.703	–
Proposed	30	0.297	0.879	0.227	0.702	0.751
w/o MEC-II	30	0.242	0.980	0.218	0.683	0.601
mMLDA	30	0.296	0.893	0.256	0.730	–
Proposed	50	0.240	0.905	0.257	0.681	0.604
w/o MEC-II	50	0.224	0.895	0.211	0.694	0.482
mMLDA	50	0.221	0.981	0.303	0.688	–
(B) 40 trials						
Proposed	0	0.304	0.960	0.240	0.691	<u>0.988</u>
w/o MEC-II	0	0.282	0.986	0.190	0.729	0.763
mMLDA	0	0.290	0.959	0.224	0.736	–
Proposed	30	0.303	0.978	0.193	0.698	0.829
w/o MEC-II	30	0.349	0.917	0.219	0.726	0.718
mMLDA	30	0.307	0.956	0.159	0.717	–
Proposed	50	0.316	0.922	0.199	0.718	0.668
w/o MEC-II	50	0.258	0.937	0.210	0.726	0.639
mMLDA	50	0.297	0.989	0.123	0.687	–

Bold and underscore indicate the highest evaluation values, and bold indicates the second highest evaluation values.

a total of 12 different test-sentences, each of which comprised four words representing the four sensory-channels. The placement of objects on the table was different in each trial and differed from the placements that were used during the learning trials.

5.2.2. Results

Figure 9 shows an example of the results of the action generation task. **Figure 9A** shows the result of action generation by the robot for the sentence “grasp front red ball.” **Figure 9B** shows the result of action generation by the robot for the sentence “reach right red cup.” **Figure 9C** shows the result of action generation by the robot for the sentence “look-at left yellow cup.” The resulting OAR values of the proposed method and its w/o MEC-II were 1.000, and the OAR value of mMLDA was 0.833. As a result, the robot could select an appropriate object even in situations and sentences not used at the CSL.

5.3. Action Description Task

5.3.1. Conditions

In this experiment, the robot performs the action and speaks the sentence regarding this action. The robot uses the learning results of the CSL task in Section 5.1. The experimental procedure is the same as in Section 4.4.1. We use the F1-measure and ACC as evaluation values. We performed the action description task for a total of 10 actions. The placement of objects on the table was different for each trial and differed from those used during the learning trials. The robot generates a sentence of four words representing the four sensory-channels. The word order in the sentence is fixed as $F_d = (a, p, c, o)$.

5.3.2. Results

Table 5 shows F1-measure and ACC values of action description task using the learning results under the different conditions. The

TABLE 5 | Experimental results in 25 and 40 trials.

Method	Trials	ROW	F1	ACC
Proposed	25	0	<u>0.575</u>	<u>0.650</u>
w/o MEC-II	25	0	0.558	0.640
mMLDA	25	0	0.406	0.558
Proposed	40	0	0.618	<u>0.720</u>
w/o MEC-II	40	0	<u>0.654</u>	0.698
mMLDA	40	0	0.509	0.645

Bold and underscore indicate the highest evaluation values, and bold indicates the second highest evaluation values.

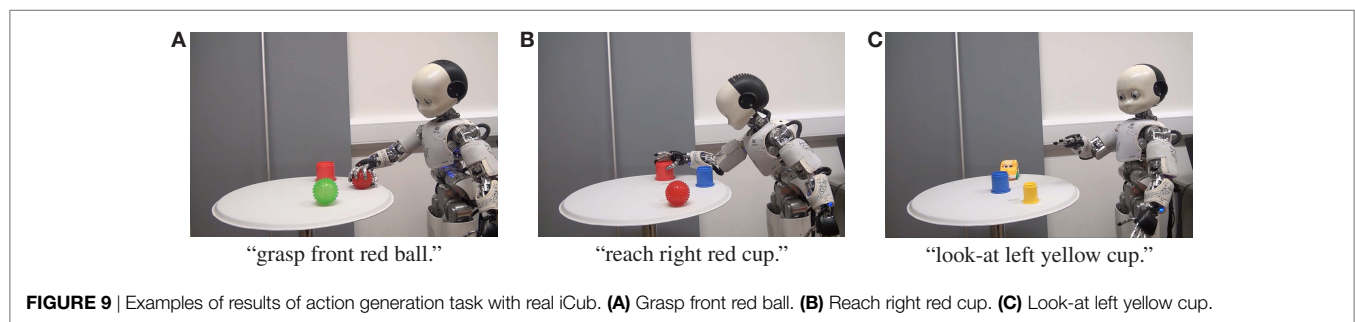


FIGURE 9 | Examples of results of action generation task with real iCub. (A) Grasp front red ball. (B) Reach right red cup. (C) Look-at left yellow cup.

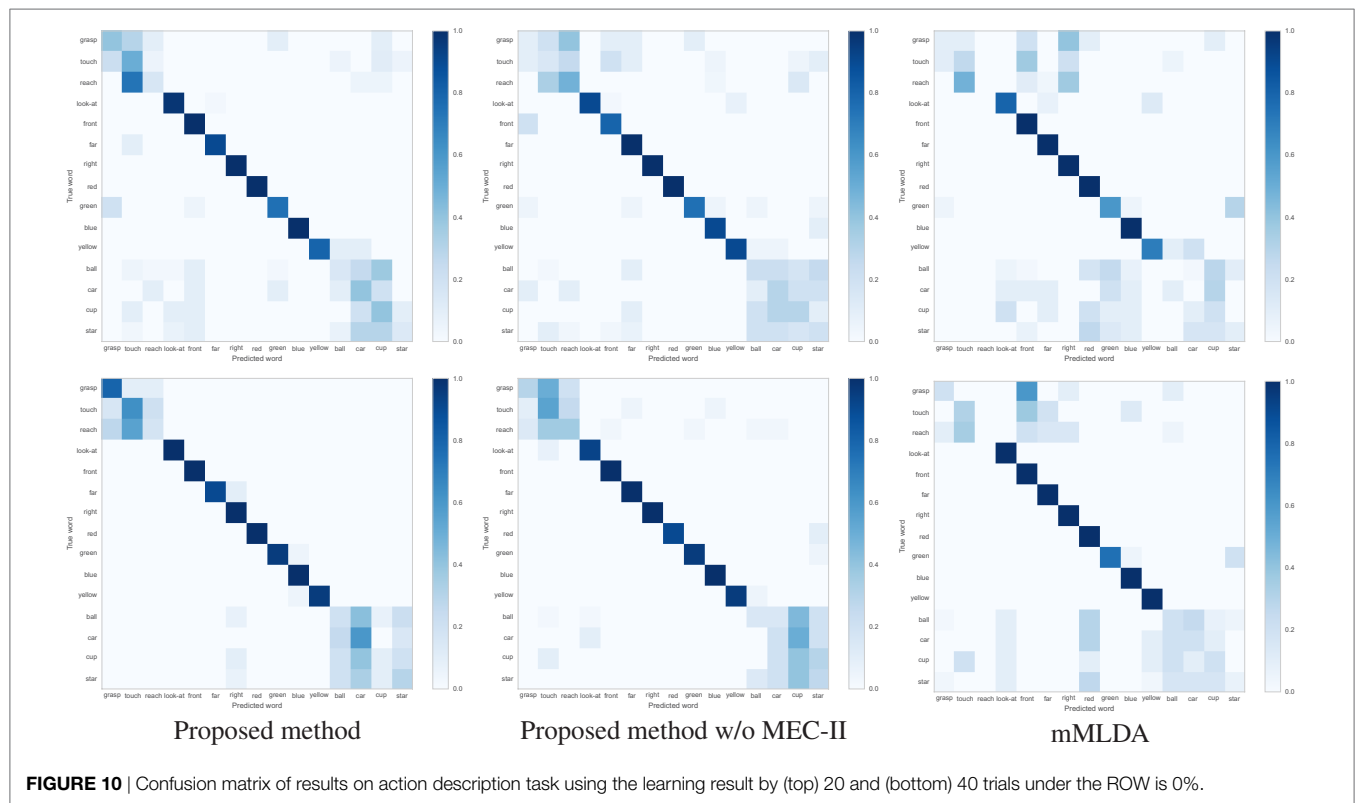


FIGURE 10 | Confusion matrix of results on action description task using the learning result by (top) 20 and (bottom) 40 trials under the ROW is 0%.

proposed method showed the higher evaluation values than other methods. **Figure 10** shows confusion matrices between predicted words and true words using the learning results by 20 and 40 trials. In the action category, there was a tendency to confuse words “reach” and “touch” similar to simulation. The major difference in the result for each method was found in the words of action and object categories. Even if the accuracy of categorization is low as in action and object categories and the categories include uncertainty, the robot could describe the action more correctly if the correspondence between the word and the sensory-channel was performed more properly.

6. CONCLUSION

In this paper, we have proposed a Bayesian generative model that can estimate multiple categories and the relationships between words and multiple sensory-channels. We performed experiments of cross-situational learning using the simulator and real iCub robot in complex situations. The experimental results showed that it is possible for a robot to learn the combination between a sensory-channel and a word from their co-occurrence in complex situations. The proposed method could learn word meanings from uncertain sentences, i.e., the sentence including four words or less with a changing order. In comparative experiments, we showed that the mutual exclusivity constraint is effective in the lexical acquisition by CSL. In addition, we performed experiments of action generation task and action description task by the robot that learned word meanings. The action generation task confirmed that the robot could also select an object successfully and generate an action even for situations other than those it

encountered during the learning scenario. The action description task confirmed that the robot was able to use the learned word meanings to explain the current situation.

The accuracy of the categorization of objects and actions tended to be lower than those of the color and position categories. In this paper, we used GMM for the categorization of each sensory-channel. MLDA achieved highly accurate object categorization by integrating multimodal information (Nakamura et al., 2011a). The accuracy of object categorization can be improved by using MLDA instead of GMM, i.e., by increasing the number of sensory-channels for the object categories. In the action categorization, the robot confused “reach” and “touch,” because these are similar actions. However, the robot is able to classify diverse actions more accurately. In addition, we used static features as action information. The accuracy can be improved by segmenting the time-series data of the actions by using a method based on the hidden Markov model (HMM) (Sugiura et al., 2011; Nakamura et al., 2016).

In this study, we performed the action generation task by a sentence including four words corresponding to the four sensory-channels. However, action instruction also presented cases in which an uttered sentence contains uncertainty. In future, we plan to investigate what kind of action the robot performs based on uncertain utterances, such as when the number of words is fewer than four, when the same objects exist on the table, and when the sentence contains the wrong word. If the robot can learn the word meanings more accurately, the robot would be able to perform an action successfully even from an utterance including uncertainty. Detailed and quantitative evaluation of such advanced action generation tasks is a subject for future work.

Other factors we aim to address in future studies are grammatical information, which was not considered in the present study, and sentences containing five words or more. We showed that the robot could accurately learn word meanings without considering grammar in the scenario of this study. However, it is important to include even more complicated situations with more natural sentences such as “grasp the red box beside the green cup.” More complicated sentences would require us to consider a method that takes the grammatical structure into account. We, therefore, aim to extend the proposed method to more complicated situations and natural sentences. Attamimi et al. (2016) used HMM for the estimation of transition probabilities between words based on concepts, as a post-processing step of mMLDA. However, they were unable to use grammatical information to learn the relationships between words and categories. Hinaut et al. (2014) proposed a method based on recurrent neural networks for learning grammatical constructions by interacting with humans, which is related to the study of an autobiographical memory reasoning system (Pointeau et al., 2014). Integrating such methods with the proposed method may be effective for action generation and action description using more complicated sentences.

In this paper, we focused on mutual exclusivity of words indicating categories in language acquisition. However, there are hierarchies of categories, e.g., ball and doll belong to the toy category. Griffiths et al. (2003) proposed a hierarchical LDA (hLDA), which is a hierarchical clustering method based on a Bayesian generative model, and it was applied to objects (Ando et al., 2013) and places (Hagiwara et al., 2016). We consider the possibility of applying hLDA to the proposed method for hierarchical categorization of sensory-channels.

For future work, we also plan to demonstrate the effectiveness of the proposed method by employing a long-term experiment that uses a larger number of objects. We believe that the robot can learn more categories and word meanings based on more experience.

REFERENCES

- Aly, A., Taniguchi, A., and Taniguchi, T. (2017). “A generative framework for multimodal learning of spatial concepts and object categories: an unsupervised part-of-speech tagging and 3d visual perception based approach,” in *Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*.
- Ando, Y., Nakamura, T., Araki, T., and Nagai, T. (2013). “Formation of hierarchical object concept using hierarchical latent dirichlet allocation,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Tokyo: IEEE), 2272–2279.
- Attamimi, M., Ando, Y., Nakamura, T., Nagai, T., Mochihashi, D., Kobayashi, I., et al. (2016). Learning word meanings and grammar for verbalization of daily life activities using multilayered multimodal latent Dirichlet allocation and Bayesian hidden Markov models. *Adv. Robot.* 30, 806–824. doi:10.1080/01691864.2016.1172507
- Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* 19, 263–311.
- Cangelosi, A., and Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots. Intelligent Robotics and Autonomous Agents Series*. MIT Press.
- Celikkanat, H., Orhan, G., Pugeault, N., Guerin, F., Sahin, E., and Kalkan, S. (2014). “Learning and using context on a humanoid robot using latent Dirichlet allocation,” in *Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Genoa.
- Chen, Y., Bordes, J.-B., and Filliat, D. (2016). “An experimental comparison between nmf and lda for active cross-situational object-word learning,” in *Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Paris.
- Fontanari, J. F., Tikhonoff, V., Cangelosi, A., Ilin, R., and Perlovsky, L. I. (2009). Cross-situational learning of object-word mapping using neural modeling fields. *Neural Netw.* 22, 579–585. doi:10.1016/j.neunet.2009.06.010
- Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *Ann. Appl. Stat.* 5, 1020–1056. doi:10.1214/10-AOAS395
- Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2007). “A bayesian framework for cross-situational word-learning,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, Vancouver, 457–464.
- Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers’ referential intentions to model early cross-situational word learning. *Psychol. Sci.* 20, 578–585. doi:10.1111/j.1467-9280.2009.02335.x
- Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: exploring the effects of context. *Cognition* 112, 21–54. doi:10.1016/j.cognition.2009.03.008
- Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., and Blei, D. M. (2003). “Hierarchical topic models and the nested Chinese restaurant process,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, Vancouver, 17–24.
- Hagiwara, Y., Masakazu, I., and Taniguchi, T. (2016). “Place concept learning by hMLDA based on position and vision information,” in *Proceedings of the 13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems (IFAC HMS)*, Kyoto.

In addition, as a further extension of the proposed method, we intend increasing the types of sensory-channels, adding a positional relationship between objects, and identifying words that are not related to sensory-channels. For example, Aly et al. (2017) learned object categories and spatial prepositions by using a model similar to the proposed model. It would be possible to merge the proposed method with this model in the theoretical framework of the Bayesian generative model. This combined model is expected to enable the robot to learn many different word meanings from situations more complicated than the scenario in this study.

AUTHOR CONTRIBUTIONS

AT, TT, and AC conceived, designed the research, and wrote the paper. AT performed the experiment and analyzed the data.

FUNDING

This work was partially supported by JST CREST and JSPS KAKENHI Grant Number JP17J07842. This material is partially based upon work supported by the Air Force Office of Scientific Research grant number: A9550-15-1-0025, Air Force Materiel Command, USAF under Award No. FA9550-15-1-0025, and by the EU H2020 Marie Skłodowska-Curie European Industrial Doctorate APRIL (674868).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at <https://www.frontiersin.org/articles/10.3389/fnbot.2017.00066/full#supplementary-material>.

VIDEO S1 | Cross-situational learning scenario using a real iCub.

VIDEO S2 | Action generation task and action description task using a real iCub.

- Harnad, S. (1990). The symbol grounding problem. *Phys. D Nonlinear Phenom.* 42, 335–346. doi:10.1016/0167-2789(90)90087-6
- Heath, S., Ball, D., and Wiles, J. (2016). Lingodroids: cross-situational learning for episodic elements. *IEEE Trans. Cogn. Dev. Syst.* 8, 3–14. doi:10.1109/TAMD.2015.2442619
- Heymann, J., Walter, O., Haeb-Umbach, R., and Raj, B. (2014). “Iterative Bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence.
- Hinaut, X., Petit, M., Pointeau, G., and Dominey, P. F. (2014). Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks. *Front. Neurobot.* 8, 1–17. doi:10.3389/fnbot.2014.00016
- Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735
- Hörnstein, J., Gustavsson, L., Santos-Victor, J., and Lacerda, F. (2010). “Multimodal language acquisition based on motor learning and interaction,” in *From Motor Learning to Interaction Learning in Robots*, eds O. Sigaud and J. Peters (Berlin, Heidelberg: Springer), 467–489. doi:10.1007/978-3-642-05181-4_20
- Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 193–218. doi:10.1007/BF01908075
- Imai, M., and Mazuka, R. (2007). Language-relative construal of individuation constrained by universal ontology: revisiting language universals and linguistic relativity. *Cogn. Sci.* 31, 385–413. doi:10.1080/15326900701326436
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). Caffe: convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, Nevada, 1097–1105.
- Markman, E. M., and Hutchinson, J. E. (1984). Children’s sensitivity to constraints on word meaning: taxonomic versus thematic relations. *Cogn. Psychol.* 16, 1–27. doi:10.1016/0010-0285(84)90002-1
- Markman, E. M., and Wachtel, G. F. (1988). Children’s use of mutual exclusivity to constrain the meanings of words. *Cogn. Psychol.* 20, 121–157. doi:10.1016/0010-0285(88)90017-5
- Matuszek, C., Fitzgerald, N., Zettlemoyer, L., Bo, L., and Fox, D. (2012). “A joint model of language and perception for grounded attribute learning,” in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, Edinburgh, 1671–1678.
- Morse, A. F., Belpaeme, T., Cangelosi, A., and Smith, L. B. (2010). “Thinking with your body: modelling spatial biases in categorization using a real humanoid robot,” in *Proceedings of the Annual Meeting of the Cognitive Science Society* (Portland, USA), 1362–1368.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.
- Nakamura, T., Araki, T., Nagai, T., and Iwahashi, N. (2011a). Grounding of word meanings in latent Dirichlet allocation-based multimodal concepts. *Adv. Robot.* 25, 2189–2206. doi:10.1163/016918611X595035
- Nakamura, T., Nagai, T., and Iwahashi, N. (2011b). “Multimodal categorization by hierarchical Dirichlet process,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (San Francisco, CA: IEEE), 1520–1525.
- Nakamura, T., Iwata, K., Nagai, T., Mochihashi, D., Kobayashi, I., Asoh, H., et al. (2016). “Continuous motion segmentation based on reference point dependent GP-HSMM,” in *Proceedings of the IROS Workshop on Machine Learning Methods for High-Level Cognitive Capabilities in Robotics*, Daejeon.
- Pointeau, G., Petit, M., and Dominey, P. F. (2014). Successive developmental levels of autobiographical memory for learning through social interaction. *IEEE Trans. Auton. Ment. Dev.* 6, 200–212. doi:10.1109/TAMD.2014.2307342
- Qu, S., and Chai, J. Y. (2008). “Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, 244–253.
- Qu, S., and Chai, J. Y. (2010). Context-based word acquisition for situated dialogue in a virtual world. *J. Artif. Intell. Res.* 37, 247–278. doi:10.1613/jair.2912
- Roy, D., and Pentland, A. (2002). Learning words from sights and sounds: a computational model. *Cogn. Sci.* 26, 113–146. doi:10.1207/s15516709cog2601_4
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Stat. Sin.* 4, 639–650.
- Smith, K., Smith, A. D., and Blythe, R. A. (2011). Cross-situational learning: an experimental study of word-learning mechanisms. *Cogn. Sci.* 35, 480–498. doi:10.1111/j.1551-6709.2010.01158.x
- Smith, L. B., and Samuelson, L. (2010). “Objects in space and mind: from reaching to words,” in *Spatial Foundations of Cognition and Language: Thinking Through Space*, eds K. S. Mix, L. B. Smith, and M. Gasser (Oxford: Oxford University Press), 188–207. doi:10.1093/acprof:oso/9780199553242.001.0001
- Spranger, M. (2015). “Incremental grounded language learning in robot-robot interactions-examples from spatial language,” in *Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Providence, RI: IEEE), 196–201.
- Spranger, M., and Steels, L. (2015). “Co-acquisition of syntax and semantics: an investigation in spatial language,” in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, Buenos Aires, 1909–1915.
- Steels, L., and Hild, M. (2012). *Language Grounding in Robots*. New York: Springer Science & Business Media.
- Stramandinoli, F., Marocco, D., and Cangelosi, A. (2017). Making sense of words: a robotic model for language abstraction. *Auton. Robots.* 41, 367–383. doi:10.1007/s10514-016-9587-8
- Sugiura, K., Iwahashi, N., Kashioka, H., and Nakamura, S. (2011). Learning, generation and recognition of motions by reference-point-dependent probabilistic models. *Adv. Robot.* 25, 825–848. doi:10.1163/016918611X563328
- Taniguchi, A., Taniguchi, T., and Inamura, T. (2016a). “Simultaneous estimation of self-position and word from noisy utterances and sensory information,” in *Proceedings of the 13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems (IFAC HMS)*, Kyoto.
- Taniguchi, A., Taniguchi, T., and Inamura, T. (2016b). Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences. *IEEE Trans. Cogn. Dev. Syst.* 8, 285–297. doi:10.1109/TCDS.2016.2565542
- Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T., and Asoh, H. (2016c). Symbol emergence in robotics: a survey. *Adv. Robot.* 30, 706–728. doi:10.1080/01691864.2016.1164622
- Taniguchi, T., Nakashima, R., Liu, H., and Nagasaka, S. (2016d). Double articulation analyzer with deep sparse autoencoder for unsupervised word discovery from speech signals. *Adv. Robot.* 30, 770–783. doi:10.1080/01691864.2016.1159981
- Tikhonoff, V., Cangelosi, A., Fitzpatrick, P., Metta, G., Natale, L., and Nori, F. (2008). “An open-source simulator for cognitive robotics research: the prototype of the icub humanoid robot simulator,” in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems* (Gaithersburg: ACM), 57–61.
- Tomasello, M., and Farrar, M. J. (1986). Joint attention and early language. *Child Dev.* 57, 1454–1463. doi:10.2307/1130423
- Twomey, K. E., Morse, A. F., Cangelosi, A., and Horst, J. S. (2016). Children’s referent selection and word learning: insights from a developmental robotic system. *Interact. Stud.* 17, 93–119. doi:10.1075/is.17.1.05two
- Yamada, T., Murata, S., Arie, H., and Ogata, T. (2015). “Attractor representations of language-behavior structure in a recurrent neural network for human-robot interaction,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Hamburg: IEEE), 4179–4184.
- Yamada, T., Murata, S., Arie, H., and Ogata, T. (2016). Dynamical integration of language and behavior in a recurrent neural network for human-robot interaction. *Front. Neurobot.* 10:5. doi:10.3389/fnbot.2016.00005
- Zhong, J., Cangelosi, A., and Ogata, T. (2017). “Toward abstraction from multimodal data: empirical studies on multiple time-scale recurrent models,” in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Taniguchi, Taniguchi and Cangelosi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.